II: Maps, Markers, and the Five-Year Goals

Physical Mapping

Maynard Olson: The idea of the Genome Project is fundamentally a strong one, but when first broached, it was an idea whose time had *almost* come. Now, five years after the first serious proposals, we're actually beginning to do something. The early proponents could be called either visionaries or cranks, depending on how generous you are. Like Jules Verne and H. G. Wells, who had clear visions of space travel but no ideas of how to implement it, the early proponents of the Genome Project had the right instincts, but they were technically naive. Their predictions—that mapping the genome would take six months and that rough sequencing of a chromosome would take a similarly brief time—were simply nonsense. Mapping and sequencing the human genome is going to be expensive, and it's going to take a long time.

Bob Moyzis: In the last five years, however, we've had some technological breakthroughs that make the Genome Project feasible, especially the first step of constructing physical maps for the whole genome. When people first started talking about this project, most of them were unaware that Maynard was working on a method to clone very large pieces of DNA in YACs [yeast artificial chromosomes]. That new cloning method has now become the mainstay of a physical-mapping effort. Without YACs, we would have been stuck with little pieces of the physical map and no way to put them together. To use an analogy, we would have had an interstate highway that was interrupted every mile or so by a stretch of dirt road or no road at all. That's better than nothing, but it's not as useful or efficient as a continuous highway.

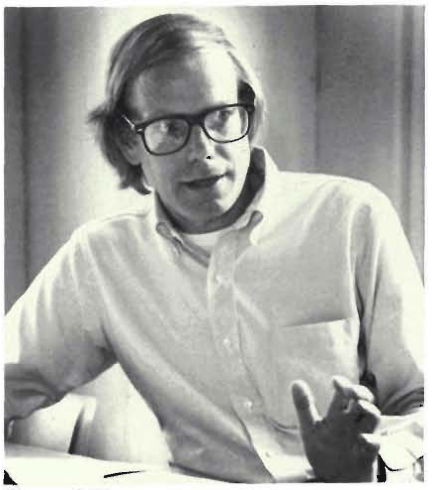
David Cox: We should point out that the physical maps we're trying to construct are not just ordinary maps of landmarks and distances. Rather, each is a reconstruction of the DNA molecule in a chromosome as a set of cloned DNA fragments. The maps are made by isolating many copies of the whole genome, cutting the DNA molecules into relatively small pieces, and cloning the pieces. Then the challenge is to figure out how to hook those pieces together in the order in which they appear along each of the twenty-four different chromosomes in the human genome.

The mapping process is much like putting together the pieces of a one-dimensional jigsaw puzzle. In the case of a DNA puzzle, the pieces are cut so that they have overlaps with neighboring pieces, and the problem is to find the overlaps and thereby order the pieces. If you succeed in putting the puzzle together, you know the exact position of each fragment relative to all the other fragments, so you can pick out exactly those fragments that span a region containing a gene of interest. Finally, you can then examine the fragments at the molecular level using all the standard techniques of molecular biology. [See "Physical Mapping—A One-Dimensional Jigsaw Puzzle."]

The difficulty in making a physical map is that often you get a few pieces hooked together to form a little island of the puzzle—that island is called a contig because it contains pieces of DNA that are contiguous in the genome—but then you get stuck because you can't find the overlapping pieces that would extend the island on each end. That's what Bob was referring to with his highway analogy. There are two reasons for getting stuck. First, the overlapping pieces you're looking for may have been lost in the cloning process, and second, your method for detecting overlaps may not be robust enough to find all of them. You end up with a whole bunch of little contigs, but you don't know how to put them together to form a whole DNA molecule. In other words, there are gaps in the puzzle.

Obviously, if you can start with larger pieces, larger cloned DNA fragments, you wind up with much longer contigs and many fewer gaps in the puzzle. That's why YACs were a breakthrough for mapping. YAC clones contain human DNA inserts that are, on average, about 300,000 base pairs in length, which is longer by a factor of 8 to 10 than the longest inserts in the clones used in earlier mapping projects. So we gained a factor of at least 10 in the speed of mapping.

Bob Moyzis: We gained speed, but more important, we gained the ability to build long contigs spanning several million base pairs of DNA. Contig maps had been constructed before in the search for disease genes, but only with great



Maynard Olson

Like Jules Verne and H.G. Wells, who had clear visions of space travel but no ideas of how to implement it, the early proponents of the Genome Project had the right instincts, but they were technically naive. Their predictions that mapping the genome would take six months, and that rough sequencing would take a similarly brief time, were simply nonsense.

difficulty and only for relatively small stretches of a chromosome known to contain interesting genes. Those maps were built with lambda-phage or cosmid clones, which carry DNA inserts of 15,000 base pairs and 40,000 base pairs, respectively. Those numbers sound large, but to cover a whole chromosome 100 million base pairs in length would require about 7000 lambda-phage clones or 2500 cosmid clones. Furthermore, constructing a physical map of overlapping cloned segments requires at least five times those numbers to ensure adequate overlaps. So neither lambda-phage clones nor cosmid clones are ideal for mapping a whole chromosome.

But the real problem was already mentioned by David Cox. When we are constructing a contig, we often can't find the clones that extend the contig. In fact, a contig map made from cosmid clones typically consists of separate contigs whose average length is about 100,000 base pairs. Until YACs came along, that was the state of the art. We could construct a high-resolution physical map—a contig—for a region 100,000 base pairs in length, and if we wanted to, we could subclone the individual clones in the contig and apply standard sequencing techniques to go down to the highest-resolution map, which is the DNA sequence itself.

In addition we could make a low-resolution map of a chromosome using a technique called in-situ hybridization to map the DNA markers present on a linkage map of the chromosome onto the chromosome itself. The markers are separated, on average, by millions of base pairs. So knowing that a disease gene was flanked by two markers didn't necessarily lead to assigning the gene to a single contig because the available contigs were shorter—by a factor of 10 to 100—than the distance between the markers. We needed a source of longer continuous pieces of DNA so that we could build contigs as long as the distance between the markers.

In the last few years the gap between a hundred thousand base pairs and several million base pairs has been filled in by two techniques. One, called pulsed-field gel electrophoresis, gives us fragments with an average length of about a million base pairs. That technique is useful but less so than first imagined because it does not always yield the same set of fragments, and further, it does not give us the DNA in a cloned form.

YAC cloning, in contrast, is a real breakthrough. It makes possible the construction of contigs a few million base pairs long. And such long continuous cloned regions bridge the gap between the high-resolution cosmid contigs and the low-resolution marker maps. We need *both* high connectivity, supplied by YACs, *and* high resolution, supplied by cosmids.

Maynard Olson: That's exactly why our five-year goal for physical mapping calls for the construction of contigs that span at least 2 million base pairs. Physical maps with such long-range continuity are essential. They are useful as navigational tools because once we find that a gene is flanked by two markers on a linkage map, we will be able to find a single contig containing both those markers and thus the gene that lies between them.

But maps are not merely navigational tools. They also provide a means of correlating many types of data. For example, we use maps to locate mountains, rivers, and city and state boundaries, but we also use maps to plot population density, average rainfall, climate changes, earthquake activity,

and so on. And once we plot those data on a map, we start to see relationships.

Cytogeneticists, doctors, and molecular biologists are all making observations on the genomes of individuals on a daily basis, but without a map we have no way of correlating those data with other information about the genome. Once we have a continuous contig map, those data will become important. We'll be able to locate the exact site of, say, a chromosomal translocation, insertion, or deletion or a new DNA marker and to correlate that information with other facts about that region of the genome. A number of labs have already constructed YAC contigs spanning several million bases, and we can expect that kind of success to continue.

David Galas: The physical-mapping projects at Los Alamos and Livermore started early, and the people there began with the much smaller cosmid clones. Over the last few years they have built 100,000-base-pair contigs, which together compose a large fraction of chromosomes 16 and 19. And now they are using YACs to bridge the gaps between the short contigs and build the long contigs that we need. In fact, the whole community is learning to use YACs for physical mapping even though they do pose some problems. In particular, many YAC clones are chimeras. That is, they contain pieces of DNA from two or more locations in the human genome. Right now those chimeric clones are a tremendous headache for the mappers.

David Cox: It's like having a fifty-piece jigsaw puzzle in which ten of the pieces are from another puzzle but you don't know which ten.

David Galas: Presumably, chimeric clones are produced by recombination between the human DNA inserts in two YACs that have entered the same yeast cell. The large amount of repetitive DNA in human DNA makes it a wonderful target for recombination in yeast. The best data about chimeric YACs come from Maynard's group at Washington University. Remember, YACs are relatively new, and chimeras were found among the clones propagated in *E. coli* too, until we came up with a strain of *E. coli* that was recombination-free. In the meantime it's very important that the mapping efforts continue despite the difficulties.

Bob Moyzis: In the last year our efforts at Los Alamos have effectively eliminated the YAC chimera problem. Starting with many copies of a single chromosome isolated by the specialized technique of flow sorting, Mary Kay McCormick has generated chromosome-specific YAC libraries of human chromosomes 16 and 21 that appear to be relatively free of chimeric clones.

The trick was to expose the yeast cells to extremely dilute YAC solutions so that the probability of two YACs entering a single yeast cell was greatly reduced and also to greatly reduce the number of recombinagenic broken DNA ends in the mixture. Clearly that's one approach to generating chimera-free YAC libraries. But other approaches need to be pursued as well, and our Russian collaborators, Vladimir Larionov and Natasha Kouprina, have had encouraging results using yeast mutants deficient in recombination.

Nancy Wexler: Perhaps we should discuss what motivates individuals to generate physical maps. It's an extremely difficult activity.

Bob Moyzis: I'm among those who are interested in the long-range order of the chromosome and therefore find the

Bob Moyzis

The structural organization of human DNA holds the key to understanding function . . . . Those who are primarily interested in finding disease genes look upon physical mapping as spending time in the barrel. They are very impatient to get back to studying some interesting disease gene.

Maynard Olson

John Sulston, the originator of the nematode-mapping project . . . is famous for working out the complete embryonic lineage of the nematode. He literally spent several years in a closet looking through a microscope at those tiny, transparent worms and watching all the cell divisions that occur as the fertilized egg develops into the mature organism.

mapping effort intrinsically interesting. As I mentioned earlier, I firmly believe that the structural organization of human DNA holds the key to understanding function. I love solving structural problems. DNA is a beautiful molecule. I *see* the DNA in every living thing. I'm just amazed by nature and driven to understand how it works. But others, those who are primarily interested in finding disease genes, look upon physical mapping as spending time in the barrel. They are very impatient to get back to studying some interesting disease gene.

Maynard Olson: Those people are never going to get very much mapping done. Mapping is a complex activity, and only people obsessed with the task itself—those who don't sleep at night when they bump up against new obstacles—will see the job to completion.

Bob Moyzis: There's a not-so-subtle conflict between the mapping effort and the traditional interests of the human-genetics community. As I mentioned earlier, that community did not initiate the Human Genome Project, and I don't sense much interest on their part in global physical mapping.

Once we have mapped the regions containing known disease genes, there may be a strong push to focus in on those genes and abandon the mapping effort. For individuals like Nancy, who have dedicated most of their careers to isolating a single disease gene in the hope of finding a cure, such a focus is appropriate and commendable. But it is not the Human Genome Project.

Unfortunately, I don't see that there are very many Maynard Olsons out there who are interested in getting a complete physical map for its own sake. Maynard pioneered the physical mapping of the baker's yeast genome [*Saccharomyces cerevisiae*], which is now just about complete. His work, as well as that of John Sulston on the nematode [*Caenorhabolitis elegans*], provided the models for how to go about making long-range physical maps. Maynard, perhaps you'd like to tell us a bit more about the motivations for making those maps.

Maynard Olson: In line with Norton's comments on his early work in bacterial genetics [see Part I of this discussion], the early efforts to map the genomes of yeast and the nematode illustrate the way in which science lurches forward. Both projects began roughly ten years ago and grew from entirely different motivations.

John Sulston, the originator of the nematode-mapping project, is a consummate biologist and, by his own characterization, a puzzle-solver. John is famous for working out the complete embryonic lineage of the nematode. He spent several years in a closet looking through a microscope at those tiny, transparent worms and watching all the cell divisions that occur as the fertilized egg develops into the mature organism. He documented the complete family tree leading from a single cell to a differentiated, multicellular organism with muscle and brain—or at least neurons—and so forth. A mature worm has a total of 959 somatic cells—cells that make up the body parts as opposed to those that produce eggs or sperm—and each worm produces those 959 cells by the same series of orderly cell divisions.

With that lineage in hand, people can, for example, use lasers to destroy a particular cell in a particular branch of the lineage and see whether other cells move in to take over the functions of the dead cell and its would-be progeny or whether

the loss just causes a gap in the mature animal. John was very strong on the infrastructure development that we were talking about earlier. He recognized that the nematode would be a much more powerful experimental system if its cell lineage were sitting there making people think about nematode development in a different way.

John then went on to make a physical map of the nematode genome and that too was a pure infrastructure development. He had been around people like Fred Sanger, so, in a sense, he had grown up at the knees of the masters, but he had never done much work with DNA. Nonetheless, he understood that a physical map of the nematode genome would make that organism an even stronger experimental system, and he went after it.

Indeed, the physical map has made the nematode an immensely more powerful experimental system. Before the completion of that map, it was extremely difficult to isolate the DNA containing a nematode gene. Typically, a mutant worm was available that exhibited a specific functional defect, for example, a particular neuron might not develop or function properly in the mutant worm. Through controlled crosses, it was inferred that the defect was caused by a single mutant gene. Further, by tracing the co-inheritance of the defect with other variable nematode traits—again through controlled crosses designed to yield maximal information about linkage to other genes—the gene was located on a high-resolution genetic-linkage map [see "Classical Linkage Mapping."]

Clearly it's a lot easier to make linkage maps for experimental organisms than for humans because, first, crosses can be controlled, and second, huge numbers of progeny are available for analysis. In the case of the nematode, a week or two of genetic-linkage mapping can often localize a gene to a region forty- to eighty thousand base pairs in length, but then that piece of DNA must somehow be isolated and cloned. Now that the nematode community has constructed a good physical map of overlapping cosmid clones and correlated it with the linkage map, the cosmids that are candidates for containing the gene of interest can be taken out of the freezer

When I first saw the basic pattern I thought, "The yeast genome has a definite physical structure, and if we could figure out the coordinates of the restriction-enzyme cleavage sites, it would be useful for genetics."

and the DNA from each cosmid can be injected into the gonads of a mutant animal. If that DNA contains the gene of interest, the defect is corrected in the resulting progeny. Then, since the DNA is in hand, the function of the gene can be pursued by the standard tools of molecular biology.

I took on the mapping of the yeast genome with a different motivation. My background is in physical chemistry, and I looked at mapping the yeast genome as a structural problem analogous, in spirit at least, to the first work on the atomic structure of proteins. I remember reading Max Perutz's description of his first good x-ray diffraction pattern from hemoglobin crystals. When he saw all those spots on the film, he realized immediately that he was seeing the structure of those proteins at atomic resolution. He had not the slightest idea of how to interpret what he saw, and it took him twenty-five years to figure out how to do so, but he was very excited when he got those first data. He was sure, even then, that it would be useful for protein chemistry to know exactly where all the atoms were in a protein.

I had a similar experience when, for the first time, I saw the DNA fragments generated by digesting the yeast genome with a restriction enzyme all separated by length on an electrophoretic gel. At that time, 1974, restriction enzymes were not available commercially. Ben Hall, the yeast geneticist with whom I was working, obtained a little tube of the enzyme *Eco*RI from another laboratory. We wasted most of it by using the wrong buffer and so forth, but eventually we were able to cut some yeast DNA into fragments. We ran the fragments out on a gel, and we got the pattern of bands that made me think about the x-ray diffraction pattern of hemoglobin. The fragments were bunched together forming thousands of bands, more than you could count, but you could clearly see that the pattern comprised discrete bands. You could even see that the patterns for different yeast strains had subtle differences.

We eventually used those variations to do yeast genetics in a way that presaged the use of RFLPs as DNA markers in human genetics. When I first saw the basic pattern I thought, "The yeast genome has a definite physical structure, and if we could figure out the coordinates of the restriction-enzyme cleavage sites, it would be useful for genetics." My geneticist colleagues thought that I was crazy, but that is because—like most biologists—they were only interested

Most scientists with no experience in carrying out a large-scale mapping project—whether they are molecular biologists or not—assume that the six thousand DNA preparations and the one thousand gels represent most of the work. In fact, [for the yeast map] the ratio of time spent on specialized analysis to the time spent on routine fingerprinting and contig construction was 10 to 1.

in research that would directly address problems of biological function.

John's effort on the physical map of the nematode was more in tune with a preoccupation with immediate biological applications, whereas my efforts were motivated more by an innate belief in the importance of understanding structure. And like John I have a basic attraction to solving technical problems. My own motive for wanting to map the human genome is simply that human DNA has an exact structure, and there is a profound lesson in that.

Bob Moyzis: Maynard, perhaps you could describe how the yeast map was constructed, since it illustrates some of the difficulties of contig construction.

Maynard Olson: The basic challenge in constructing contigs of overlapping clones is to find the overlaps. One begins with a set, or a so-called library, of thousands of anonymous cloned fragments. I say anonymous because at the outset of mapping absolutely nothing is known about the fragments. The trick is to get just enough information about each fragment to be able to detect that one fragment overlaps another.

For the yeast project we picked clones at random, and then we created a fingerprint for each clone by cutting it up with a single restriction enzyme and separating the resulting fragments by length on a gel using electrophoresis. The lengths of the restriction fragments defined the fingerprint for the clone. If two clones have many restriction fragments of similar length in common, statistical arguments tell you that those two clones have a high probability of overlapping.

This procedure yielded not only contigs of overlapping clones but also a restriction map for each yeast chromosome, a map that gives the distances between restriction-enzyme cutting sites along the chromosome. [See "Physical Mapping—A One-Dimensional Jigsaw Puzzle."] The average distance between the cutting sites on the yeast maps is approximately 2000 base pairs. The detailed physical maps are now providing a solid base for present efforts to sequence the entire yeast genome.

Bob Moyzis: That's a quick description, but you have estimated that the yeast map took 20 person-years to complete. How was that time spent?

Maynard Olson: Let's look at the routine work first. We analyzed roughly six thousand clones, obtaining a single-digest fingerprint for each. To produce most of the clones, we used lambda-phage vectors, which are derived from a widely used *E. coli* virus. The lambda clones each contained about 20,000 base pairs of yeast DNA. We also made a few hundred clones with cosmid vectors, and each of those clones contained about 40,000 base pairs of yeast DNA. Since the yeast genome contains about 15 million base pairs of DNA, our collection of clones provided nearly a tenfold sampling redundancy.

The clones were analyzed ten at a time on standard electrophoretic gels. Counting analyses that needed to be repeated and those that gave no useful data, nearly a thousand gels were run. Even though all that laboratory work was done by hand, it represents no more than 10 percent of the 20 person-years Bob mentioned. Moreover, that figure does not include one-time research and development activities such as software development and methodological research needed to come up with a workable strategy for finding overlaps and constructing contigs.

Most scientists with no experience in carrying out a large-scale mapping project—whether they are molecular biologists or not—assume that the six thousand DNA preparations and the one thousand gels represent most of the work. In fact, the ratio of time spent on specialized analysis to the time spent on routine fingerprinting and contig construction was 10 to 1.

What kinds of specialized analyses were needed? Significant effort went into tracking down errors and inconsistencies in the data. We had a data set of very high quality, but still we found that 5 percent of the fingerprints were problematic because the clones were biologically anomalous—they were unstable on propagation or were of artifactual origin—and another 10 percent of the fingerprints were experimentally suspect—they were obtained from under- or over-digested DNA samples or involved mixed clones or incorrectly interpreted gel images.

Those special cases produced inconsistencies in the map, the most common being a branching contig. In other words, one contig would appear to branch into two when we attempted to accommodate all the fragments in the fingerprints of the clones in a linear order corresponding to a single contig. We used conservative criteria for recognizing overlaps, so very few of the inconsistencies resulted from placing clones in the wrong contigs. Most often, there was simply something wrong with the fingerprint data.

The key to building correct maps from reliable clone collections is to track down all the anomalies. Altogether we had about a thousand cases requiring special attention, and that attention had to come from skilled personnel and often required new experimental effort.

In addition, after the contigs were built, special experiments—none of them particularly satisfactory—were required to orient all the contigs in the same direction and align them with the chromosomes. Contigs built from lambda-phage and cosmid clones are rarely longer than 150,000 base pairs even when sensitive overlap-detection methods are employed. Therefore, the best-case scenario for a map the size of the yeast genome involves orienting and aligning one hundred contigs.

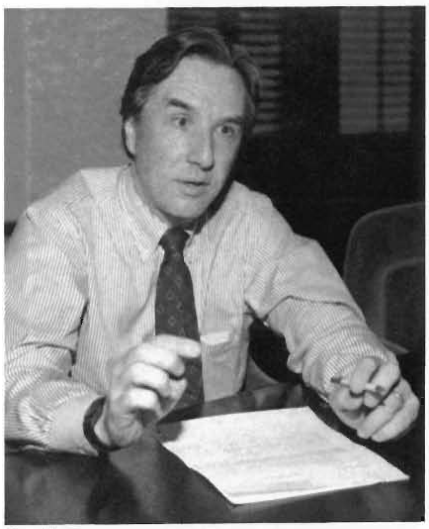
In reality, the yeast project dealt with several hundred contigs. For an average human chromosome the number would be closer to a thousand. Automation is not going to lessen the effort required to check and align large numbers of contigs, since it can only be applied to the routine activities that account for a small part of the total effort.

Bob Moyzis: Approaches similar to the one described by Maynard were used in the nematode mapping and the initial physical mapping of individual human chromosomes at Los Alamos and Livermore. At Los Alamos we realized that if more information could be rapidly obtained about each clone, then smaller overlaps could be detected, and hence, the initial mapping would progress faster. How could you obtain more information rapidly? That's where a low-resolution knowledge of the structural organization of human DNA proved useful.

Human DNA, unlike yeast and nematode DNA, is littered with multiple copies of various DNA sequences. The function of the repetitive DNA, if it has any, is unknown, leading some people to describe repetitive DNA as junk or parasitic DNA, as we mentioned earlier. Four particular sequences appear hundreds of thousands of times per

Norton Zinder

Though many people thought that the technical problems associated with large-scale mapping and sequencing would not be interesting to young people, the opposite seems to be true. Graduate students are tremendously enthusiastic about getting into this field.



David Cox

People argue about which is the right technique for mapping the genome, and some are trying to push a given technique to its limit . . . when you search that hard, you end up making errors. Somebody is not the real father, or a tube is mislabeled, so . . . the new marker is not any closer to the gene than the markers you already have. Those mistakes happened in the search for the Huntington's gene.

genome and account for 5 to 10 percent of the DNA mass. Since our low-resolution studies indicated that those sequences were essentially randomly interspersed in human DNA, we realized that the locations of the four repetitive sequences in the restriction fragments of each clone would supply the needed extra information—and we could get that information rapidly.

One can show that for cosmid-sized fragments of a single chromosome, such repetitive-sequence fingerprints are essentially unique. David Torney at Los Alamos took this basic concept and developed a mathematically rigorous algorithm to identify pairs of overlapping clones. As predicted, our mapping initially progressed four to five times faster than the mapping of the nematode, which has about as much DNA as a single human chromosome. Our work has now progressed to the closure phase. The initial 550 cosmid contigs are being linked together with YAC clones to form between 50 and 100 contigs, each with an average length of 1 million to 2 million base pairs. [See "The Mapping of Chromosome 16."]

I'd like to point out that using fingerprints—even our repetitive-sequence fingerprints—to determine whether two clones overlap is a probabilistic approach to building contigs. A more powerful approach has recently been pioneered at Maynard's laboratory. The method involves first identifying a set of so-called sequence-tagged sites [STSs]—short segments of human DNA each with a unique base sequence—and then using the polymerase chain reaction to determine which STS is present in each clone. Any two clones that both contain the same STS by definition—and without question—overlap. STSs have become the main tool for map assembly as well as a universal map language.

Maynard Olson: It was obvious at the start of the Genome Project that we needed stronger overall strategies. YACs look like a promising way of keeping the analysis modular and reducing the number of modules—the number of clones and the number of contigs—to a more manageable number. However, even when YACs are used, the effort required to bring home a reliable, well-documented map is enormous.

Sequencing the human genome is going to be an even bigger job, but I doubt that it will prove to be as qualitatively difficult as mapping. And I am confident that it will be much more amenable to automation.

Norton Zinder: The beginning and end of any science—cosmology, anatomy, or molecular biology—is based on finding out where things are relative to each other. So the genome maps are fundamental. And though many people thought that the technical problems associated with large-scale mapping and sequencing would not be interesting to young people, the opposite seems to be true. Graduate students are tremendously enthusiastic about getting into this field. They find the rest of science crowded and in a way uninteresting.

A new generation of people will come into this field without the label of being molecular biologists and with a different mind-set. They see this field as wide open, as an opportunity to get lots of new information and data. And there's nothing more satisfying to a scientist than collecting lots of data.

In the days when I was doing almost nothing but making new bacterial mutants, I'd sit down at the end of the day and fill my notebook with the fifteen new mutants I had just knocked off, feeling very, very satisfied.

Bob Moyzis: The Genome Project is in essence a very large data-gathering effort—one that requires a lot more coordination than is normally found in biology. On the other hand, efforts to map the genome are divided among many laboratories. Los Alamos is mapping chromosome 16 and parts of chromosomes 5 and 21, Livermore is mapping 19, Washington University is mapping X and 7, and so on.

One reason for dividing the project by chromosome was to preserve the structure of the biological research community. No one wanted a large, monolithic organization dictating how the information would be gathered and disseminated. Each of the genome centers is using a different set of mapping strategies, depending on the talents and expertise of the scientists involved.

It's important to emphasize that there is no *right* way to generate a physical map and that many different techniques are needed to produce and confirm the map. As long as your map is translated into the STS language, how you obtained it is not relevant. Given the nature of molecular biologists, it will be obtained in any way that works. Under its present funding structure, the NIH may adopt the strategy of building a low-resolution, one-megabase, YAC map of the entire genome at one center. This map would be used as a framework for the construction of high-resolution maps at many different laboratories.

David Cox: Right now, in our work on chromosome 4 at UCSF, we're building contigs of overlapping YAC clones using a new type of linkage analysis called radiation-hybrid mapping to determine physical distances between unique DNA markers, and we're using in-situ hybridization to order the contigs and the DNA markers.

People argue about which is the right technique for mapping the genome, and some are trying to push a given technique to its limit. But no single technique will give us a reliable map of the genome. Each one is powerful only within a certain range of resolution, and at the limits of that resolution, it becomes inefficient and inaccurate.

The moral of the story is that we need to combine many different techniques if we're going to map the genome in a reasonable time.

Here's an example of what happens when you push linkage analysis to its limits. In searching for the Huntington's-disease gene, people were able to find DNA markers flanking the gene that were deduced from linkage analysis to be about 2.5 million base pairs apart. [See "Modern Linkage Mapping."] They went on to make a physical map of overlapping clones that spanned the region between the markers. But they wanted to narrow the search even further because finding a gene in 2 million base pairs of DNA is still a tough job. So they found new candidates for flanking markers and tried to find recombination events in afflicted families that would indicate the genetic distance between the new markers and the gene. [See "Classical Linkage Mapping" for a discussion of recombination events.]

Well, not very many recombination events take place within 2 million base pairs, so you have to scan the world for all the families afflicted with Huntington's disease in the search for possible recombination events. The sad fact is that when you search that hard, you end up making errors. Somebody is not the real father, or a tube is mislabeled, so what you thought was a recombination event really is not, and the new marker is not any closer to the gene than the markers you already have. Those mistakes happened in the search for the Huntington's gene. All the workers in the field chased down a lot of garbage, and now the best we can do is to search through a region 2.5 million base pairs in length to find the gene. There's no additional recombination information at this time that allows us to find markers closer to the gene. Tomorrow there could be a new recombination event, but it's not very likely.

The moral of the story is that we need to combine many different techniques if we're going to map the genome in a reasonable time. Just as we need a series of microscope lenses with increasingly higher magnification to look not only at a whole cell but also at the small organelles within it, different mapping methods have different powers of resolution and we need all of them.

Bob Moyzis: Your point is well taken and it applies to contig building as well. For example, now that YACs are available, people argue that we should forget about cosmids because the contigs built from cosmid clones are relatively short. On the other hand, as soon as someone has a YAC contig, the first thing that person will do in order to find out more about what's in those YACs is to subclone them in cosmids or some other cloning vector. Smaller clones are much easier to work with, so cosmids will continue to play an important role in the mapping project. That is, of course, until we can directly sequence a 300,000-base-pair YAC.

Physical Mapping

a one-dimensional jigsaw puzzle

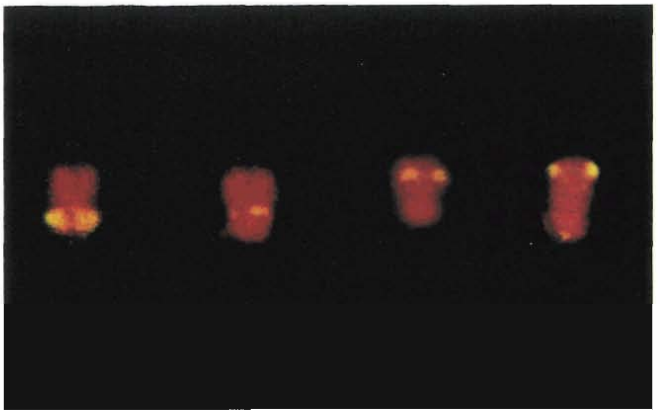
The human genome consists of forty-six double-stranded DNA molecules. Each molecule is made up, on average, of 130 million base pairs strung in a linear order between two sugar-phosphate backbones, and each is wound around proteins to form a chromosome. In order to study genes and other interesting regions of the genome at the molecular level, standard practice is to isolate the DNA and break up the long molecules into many fragments. We then make many identical copies of each fragment by cloning and pick out the clones of interest. Almost all methods for analyzing DNA at the molecular level require many copies of the fragment of interest. Therefore, cloning is essential for procedures such as finding the positions of restriction-enzyme cutting sites, determining the sequence of nucleotide bases in a particular DNA fragment, and identifying polymorphic DNA markers. However, in fragmenting the DNA molecules prior to cloning, we lose all information about the physical locations of fragments along the genome itself.

Problem: How do we find the chromosomal positions of known genes, polymorphic markers, and other cloned portions of the human genome?

Low-Resolution Physical Mapping by In-Situ Hybridization

In contrast to a linkage map, which specifies statistical distances between variable DNA markers and genes in terms of recombination fractions (see "Classical Linkage Mapping"), a physical map specifies physical distances between landmarks on the DNA molecule of each chromosome.

In-Situ Hybridization on Human Chromosome 21



Four DNA probes labeled with a fluorescent dye produce positive hybridization signals at four locations along chromosome 21. Because metaphase chromosomes are made up of two nearly identical sister chromatids, each probe produces a pair of signals.

One standard low-resolution method for finding the physical position of a cloned fragment is in-situ hybridization on metaphase chromosomes. We first find a segment within the cloned region whose base sequence occurs nowhere else in the genome. We then synthesize many copies of a single strand of that unique segment and label each copy with a fluorescent tag to make it useful as a DNA probe. A solution containing the DNA probe is then applied to a spread of chromosomes that have been arrested at metaphase and fixed to a microscope slide. (Metaphase is the phase of cell division during which chromosomes have condensed to form the wormlike shapes easily visible under a light microscope.) Under appropriate conditions the probe binds, or hybridizes, only to the chromosomal DNA with a base sequence exactly complementary to that of the probe (see "Hybridization" in "Understanding Inheritance"). The position on a metaphase chromosome where the probe has hybridized is imaged with a fluorescence microscope as a bright spot. Because DNA molecules are wound very tightly during metaphase, the resolution achieved with in-situ hybridization is low, about 3 million base pairs. In other words, the hybridization signals from two probes less than 3 million base pairs apart will overlap one another and cannot be resolved into two distinct spots. In-situ hybridization using

four cloned inserts as probes produced the bright spots on the metaphase chromosomes in the micrograph shown on the page opposite.

High-Resolution Physical Mapping by Construction of Contig Maps of Overlapping Clones

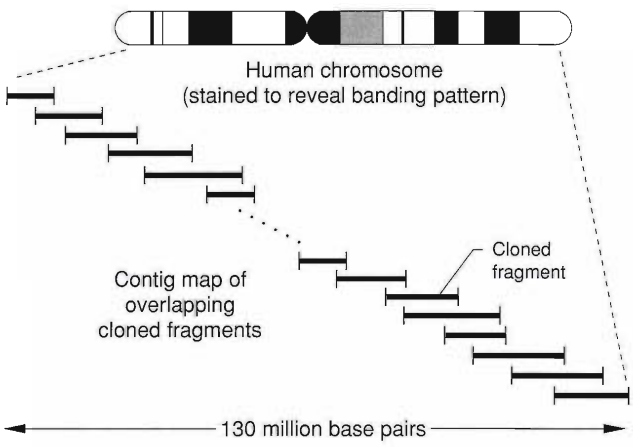
To determine the positions of genomic landmarks with much greater resolution, we can replace the chromosomes themselves with twenty-four contig maps, one for each of our twenty-two homologous chromosome pairs and one for each of our two sex chromosomes. A contig map is a set of contiguous overlapping cloned fragments that have been positioned relative to one another. In a complete contig map for a human chromosome, the cloned fragments would include all the DNA present in the chromosome and follow the same order found on the DNA molecule of the chromosome. As in any physical map, distances are measured in base pairs.

Using these contig maps, we can localize any cloned fragment or other DNA probe, again by hybridization, to a much smaller portion of the genome, namely to one of the cloned fragments in one of the maps. Moreover, we can determine the position of any DNA probe relative to all other landmarks that have been similarly localized. Once contig maps are constructed, the entire genome will be available as cloned fragments, and we will be able to use these clones to analyze any region down to the level of its base sequence.

Example: The figure at right is a schematic of a contig map for one chromosome. Right now, the top priority of the Human Genome Project is to construct a contig map for each of the twenty-four different chromosomes in the human genome. Those maps, when integrated with the corresponding genetic-linkage maps, will provide a means of finding the segments of DNA that contain disease genes (see "Modern Linkage Mapping"). The clones that make up the map also provide the material needed to sequence the human genome.

Many different strategies are being developed to make contig maps of human chromosomes. (Details of the Los Alamos effort to map a human chromosome are presented in "The Mapping of Chromosome 16.") Here we introduce the basic principles of contig-map construction.

Physical Map of a Human Chromosome



The contig map spans the single DNA molecule contained in the chromosome.

Question: How do we obtain the clones that compose the contig maps?

Answer: We prepare a collection, or library, of cloned human DNA fragments in a manner such that (1) essentially all parts of the genome are probably present in the library and (2) the human DNA fragments in the clones overlap one another. Overlaps among the cloned fragments are essential because they allow us to reconstruct the order in which the fragments appear along the genome.

Example: The figure illustrates the steps in preparing a library of cloned DNA fragments. We start by isolating the DNA from many human cells. Then we break up the DNA into a large set of overlapping fragments by partial digestion of the DNA with a restriction enzyme. A restriction enzyme digests a DNA molecule by recognizing and cleaving the molecule at every occurrence of a particular short sequence usually four to eight base pairs long. Such a site is called a restriction site and is marked on the figure by a dot. Since complete digestion would yield nonoverlapping fragments (every copy of a particular DNA molecule would be cleaved at the same places), we interrupt the digestion process before it reaches completion, thereby leaving many restriction sites intact at random locations along each molecule. (In the figure, cleavage is indicated by a vertical line through the restriction site.) Such partial digestion ensures that each resulting fragment will overlap other fragments in the set.

Next, each of these fragments is joined to a cloning vector to form a recombinant DNA molecule. A cloning vector is a small DNA molecule that, after entering a host organism (such as yeast or bacteria), is replicated by the cellular machinery of the host organism. The cloning vector shown here is a small circular DNA molecule that has been engineered to include a single cutting site for the restriction enzyme chosen to digest the sample of human DNA. Copies of the cloning vectors are cut at that site and are mixed with the human DNA fragments, and the enzyme DNA ligase is added to the mixture. The "sticky ends" of a cloning vector (which are formed by restriction-enzyme cleavage) bind to the "sticky ends" of a human DNA fragment, and the ligase catalyzes the chemical union of the sugar-phosphate backbones of the two DNAs into a recombinant DNA molecule. We then expose a population of the host organism to the recombinant DNA molecules, and, if we are lucky, each recombinant DNA molecule enters a host organism and is there replicated as the host replicates. Each host colony containing clones of a particular fragment is individually plucked and stored in a well of a 96-well microtiter dish where the cells can be grown up again and again. This library of clones provides a renewable supply of all the fragments that have survived the cloning process.

To create a contig map of a single human chromosome, many groups are starting with a chromosome-specific library of cloned fragments constructed by starting with many copies of a particular chromosome. Chromosome-specific libraries are being made by the National Laboratory Gene Library Project at Los Alamos and Livermore and are available to research groups throughout the world (see "Libraries from Flow-sorted Chromosomes").

The cloned fragments in a DNA library are "anonymous"; that is, we know nothing about them except their approximate length, which is determined by the length of the DNA insert that can be successfully incorporated into the cloning vector we have chosen. Until recently cosmids were the cloning vectors most often used for map construction. Cosmids reproduce in the bacterial host *E. coli*, and they accept DNA inserts ranging from about 25,000 to 45,000 base pairs in length. Therefore about 4000 cosmid clones could accommodate all the DNA in an average human chromosome. However, to achieve the overlaps among cloned fragments required in the construction of a contig map and to better assure that all the chromosomal DNA is represented in the clone library, the usual practice is to construct a library with up to ten times that number of cosmid clones.

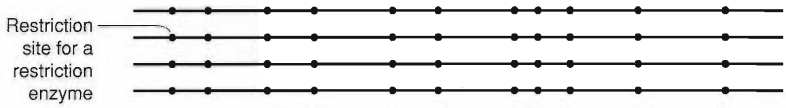
Question: How do we position the cloned DNA fragments along the DNA molecules in the genome?

Answer: Positioning cloned DNA fragments is analogous to solving a one-dimensional jigsaw puzzle, but rather than looking for interlocking pieces, we look for detectable overlaps between clones, that is, for clones that have a unique stretch of human DNA in common. Because the number of pieces in the puzzle is so large, we need a rapid method for detecting overlaps between pairs of clones. If we could sequence each clone, we could identify overlaps unambiguously, provided the overlapping region is not a sequence that repeats elsewhere in the genome. However, given the current state of sequencing technology, that approach is totally impractical.

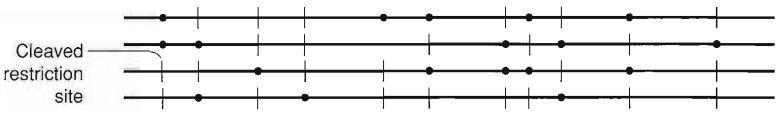
A practical and successful probabilistic method for detecting overlaps is to make a "fingerprint" of each clone (more precisely, of the human DNA insert within each clone) and compare the

Construction of a Library of Cloned DNA Fragments

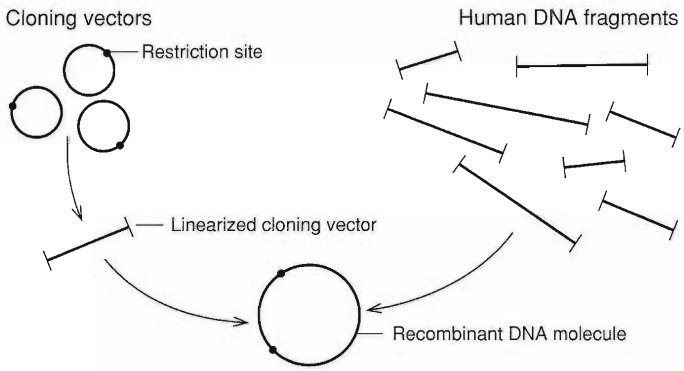
Step 1: (a) Isolate many copies of the human DNA molecule to be mapped.



(b) Partially digest the molecules with a restriction enzyme to create overlapping fragments.

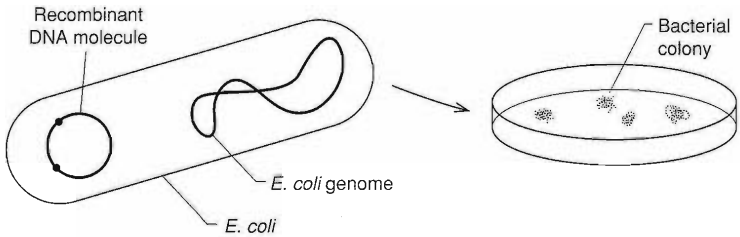


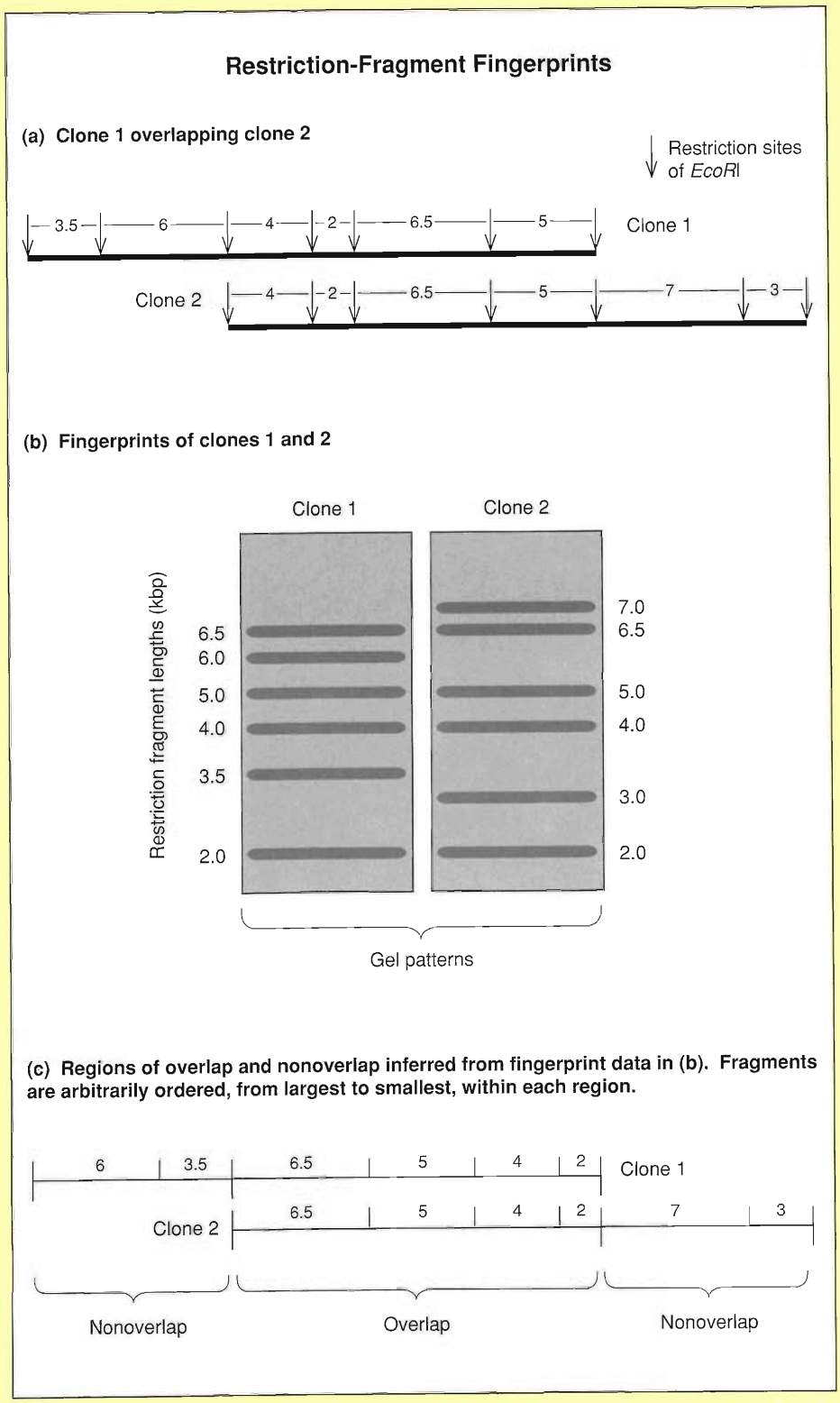
Step 2: (a) Linearize the circular cloning vectors with the restriction enzyme used in step 1b.



(b) Ligate cloning vectors and human DNA fragments to create recombinant DNA molecules.

Step 3: Facilitate the entry of recombinant DNA molecules into host cells, here the bacterium *E. coli*, and grow each host cell into an isolated colony, thereby producing many identical copies of that recombinant DNA molecule.





fingerprints. The simplest fingerprint of a cloned fragment is the one obtained by completely digesting about 10^{10} copies of the clone with a restriction enzyme and then determining the lengths of the resulting restriction fragments by gel electrophoresis. The restriction-fragment lengths determined from the gel constitute the restriction-fragment fingerprint of the clone.

Suppose we obtain restriction-fragment fingerprints of our clones by using the restriction enzyme *Eco*RI, which can cut DNA at every occurrence of the six-base-pair sequence GAATTC. Within a random sequence of the four DNA bases, any six-base-pair sequence occurs, on average, every 4^6, or about 4000, base pairs. Therefore the average length of the restriction fragments produced by *Eco*RI from a random sequence of the DNA bases is about 4000 base pairs. Now the sequence of bases in the human genome is not random, but nonetheless, the average length of the restriction fragments in the *Eco*RI fingerprints of a set of clones is about 4000 base pairs. Thus we expect that the human DNA inserts in two cosmid clones, each of which are, say, about 30,000 base pairs long, will have at least one restriction fragment in common if they overlap by more than about 15 percent.

Example: To illustrate the information content of fingerprints made by using the restriction enzyme *Eco*RI, consider two clones that are known to overlap as shown in part (a) of the figure. The cleavage sites for *Eco*RI are marked by arrows, and the distances between restriction sites are given in thousands of base pairs (kbp). Part (b) shows the restriction-fragment fingerprints obtained by completely digesting many copies of each clone with *Eco*RI. After several hours of electrophoresis, the restriction fragments of

each clone have separated into distinct bands, each band consisting of all the restriction fragments with a particular length. (The bands are made visible by staining, and each gel is calibrated with fragments of known lengths.)

The region of overlap between the two clones shown in the figure yields four restriction fragments with lengths of 4, 2, 6.5, and 5 kbp. Thus the fingerprints of the two clones have four bands in common at the gel positions corresponding to those lengths. Suppose these two fingerprints were the only information we had about the two clones shown in the figure. We might suspect that the clones overlap one another and that the overlap region included four restriction fragments with lengths of 2, 4, 5, and 6.5 kbp. We might then partition the restriction fragments into a region of overlap and two regions of nonoverlap as shown in part (c) of the figure. Note that we would have no way to impose any further ordering on the restriction fragments present in the fingerprint. Shown in (d) is a photograph of actual fingerprint data.

Question: Can we infer that two clones overlap solely on the basis of their restriction-fragment fingerprints?

Answer: Since a restriction-fragment fingerprint is, in essence, just a list of restriction-fragment lengths, it gives us no information about the order of the fragments within each clone. Also, we can't tell whether the restriction fragments of the same length in two different fingerprints are copies of the same fragment. So the fact that the fingerprints of two clones have one or more restriction-fragment lengths in common does not provide unambiguous evidence that the two clones overlap. On the other hand, by taking into account statistical properties of restriction-fragment lengths, we can estimate the likelihood of overlap given the data. David Torney of Los Alamos has developed a rigorous formulation of the likelihood calculation that takes into account the distribution of the distances between cleavage sites in the genome (the distribution of *Eco*RI cleavage sites appears to be a Poisson distribution with an average spacing of 4000 base pairs), the errors in the measurement of restriction-fragment lengths (about 1 percent), and all possible ways in which the two clones might overlap. Since the declaration of a false overlap would lead to the merging of pieces of the map that are not contiguous on the genome and since such mistakes are very time-consuming to correct, a conservative approach is to declare an overlap only if the likelihood of overlap is 90 percent or greater. Given the simple restriction-fragment fingerprints shown on the page opposite, two clones must overlap by about 50 percent to yield such high likelihoods of overlap. Thus small overlaps are typically not detected with this conservative approach. As described in "The Mapping of Chromosome 16," the Los Alamos mapping group has devised a fingerprint that includes information about the presence of repetitive DNA sequences on the restriction fragments in each fingerprint. That additional information facilitates the detection of much smaller overlaps and therefore requires the fingerprinting of fewer clones to complete the contig map.

Question: How are pairs of clones with a high likelihood of overlap assembled into contigs, sets of contiguous overlapping clones?

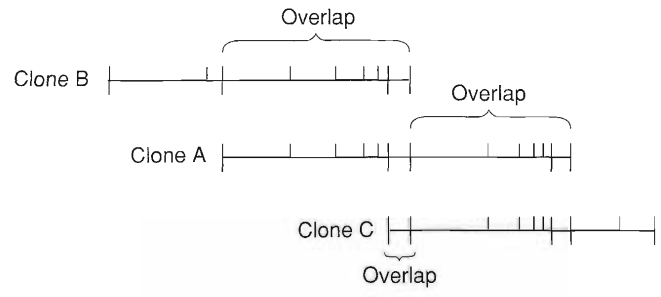
Answer: Given the uncertainties in fingerprint data, assembling pairs of overlapping clones into contigs from those data alone is a difficult computational problem. The

standard procedure is to find pairs of clones, link those pairs into groups, and then attempt to order all the restriction fragments within each group of clones in a self-consistent manner. The method is essentially an incremental approach. As each new clone is added to a contig, one tries to retain as much of the existing construction as possible even in the face of contradictory data.

A significant departure from the incremental procedure has recently been developed at Los Alamos. Map construction is treated as an optimization problem in which all available data are taken into account rather than only the data yielding high overlap probabilities. A description of this global approach to map construction is discussed in "Computation and the Human Genome Project." Here we illustrate the more standard procedure.

Example 1: Suppose that the fingerprints of clones A, B, and C reveal that clones A and B have five fragment lengths in common, A and C have six fragment lengths in common, and B and C have one fragment length in common. Furthermore, we have calculated from those data that the likelihood of A and B overlapping is 90 percent, of A and C overlapping is 95 percent, and of B and C overlapping is 10 percent. We would then assemble the three clones into a contig as shown in the figure, where some restriction fragments are placed in regions of overlap and the

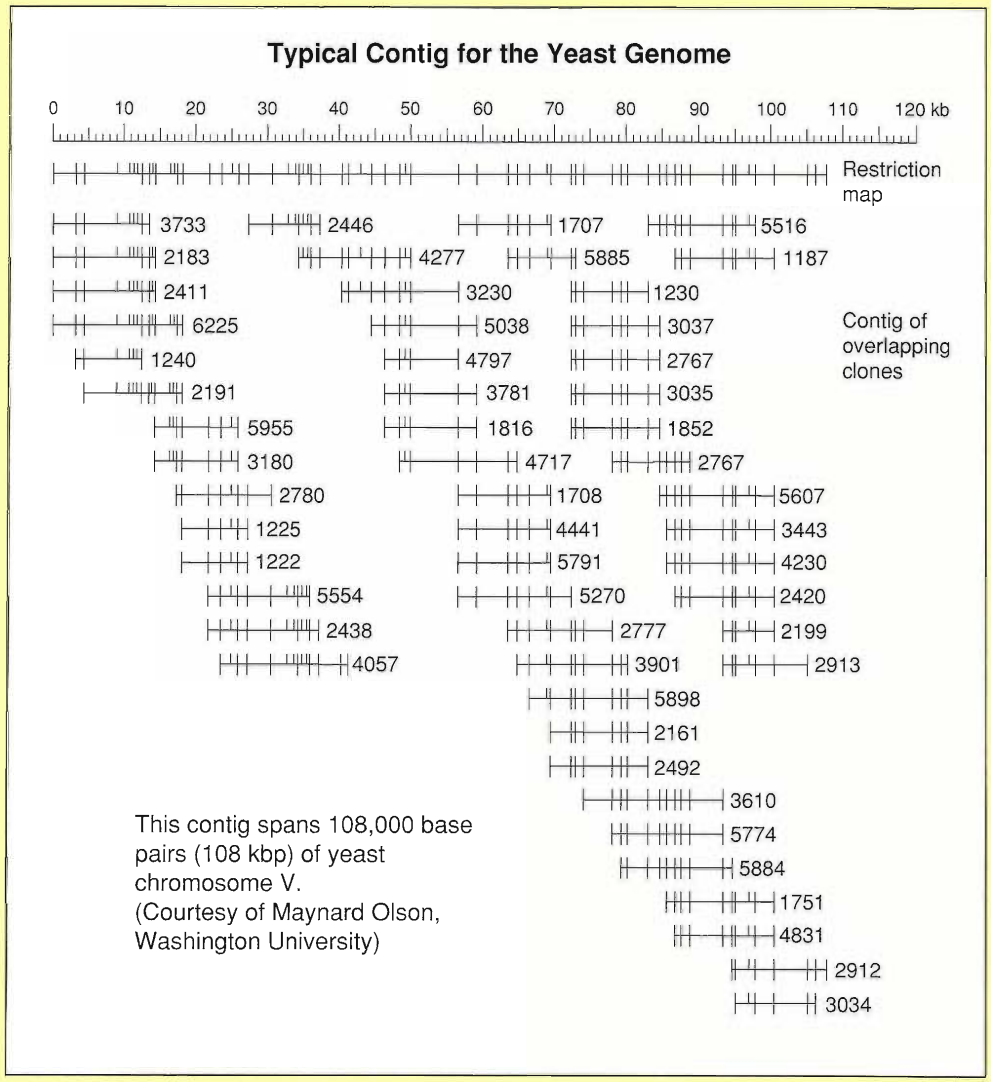
Assembly of a Contig



Likelihood analysis of fingerprint data suggests that clone A overlaps clone B and clone C and that clone B and clone C do not overlap. However, clone B and clone C do share one restriction fragment and that fragment can be placed in the overlap between clone B and C with no loss of consistency. Preliminary assignments of restriction fragments to overlap or non-overlap regions might be altered as more clones are added to the contig.

remaining ones are placed in the regions of nonoverlap. As we add other clones to the contig, we might have to revise the partitioning of the fragments into overlapping and nonoverlapping regions to construct a consistent ordering for the entire contig. Because of the uncertainties in fragment lengths and the possibility that fragments of equal length are not necessarily the *same* fragment, complicated computer algorithms are necessary to determine the most likely order of the clones in a contig. When the number of clones in a contig is much larger than the number required to span the region covered by the contig, we can order many of the restriction fragments that appear in each fingerprint and thereby help to avoid some false overlaps.

Example 2: Shown at right is a contig assembled on the basis of restriction-fragment fingerprints. The contig spans about 100,000 base pairs. Also shown is a restriction map deduced from the contig. The restriction map shows the order of and distances between restriction sites in thousands of base pairs or in kbp. The exact positions of some restriction sites (marked by the longer vertical lines that extend through the cloned fragments) have been determined by the fact that each lies at the end of one of the clones in the contig and therefore separates a region of overlap between two clones from a region of nonoverlap. Other restriction sites (marked by the shorter vertical lines) have been localized to a single overlap region but cannot be ordered further. Such sites have been arbitrarily located left to right on the contig in order of decreasing inter-site distance. This contig is representative of those used in constructing the recently completed physical map of the genome of baker's yeast (*Saccharomyces cerevisiae*). That map is, on average, eight clones deep. That is, any region is present in, on average, eight clones. Such great redundancy provided information about the order of a large fraction of the restriction sites and greatly reduced the chance of a false overlap.



Question: Do the disconnected contigs assembled by fingerprinting randomly selected clones steadily increase in length until they become connected?

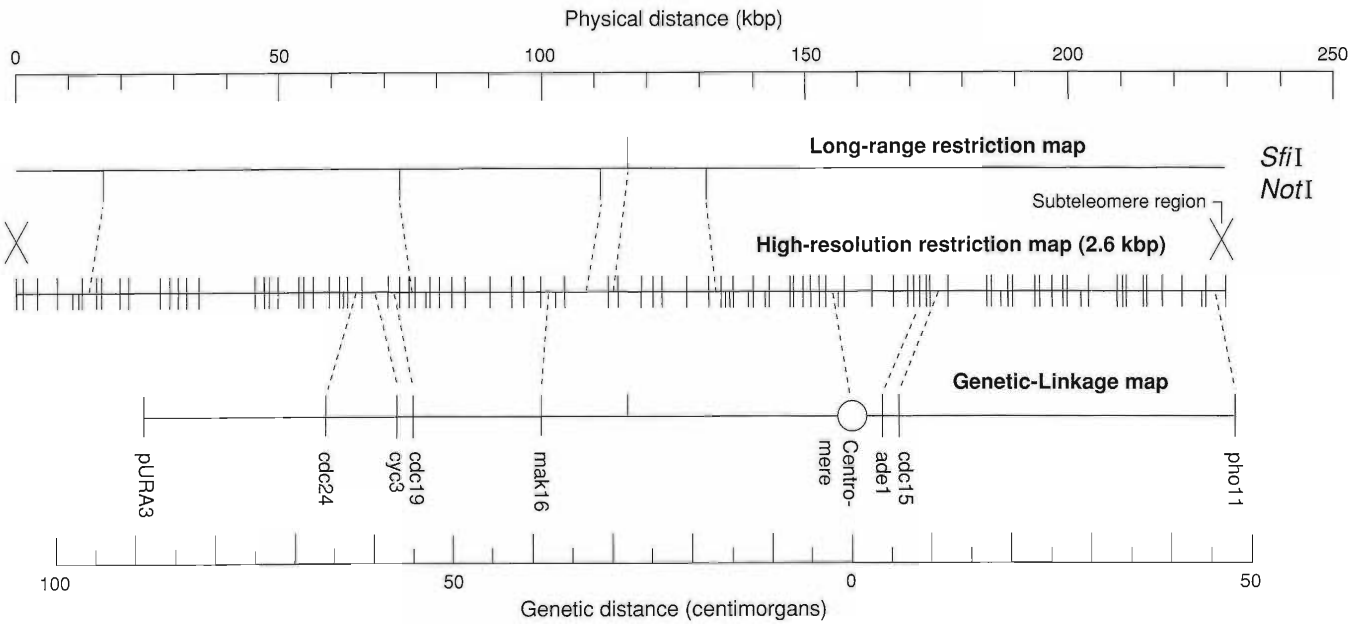
Answer: No. In a random fingerprinting strategy, both the numbers and sizes of the contigs grow fairly rapidly at first, but the rates of growth decrease after the existing contigs cover about two-thirds of the region to be mapped. The decrease in growth rate is due to the increasing probability that a randomly selected clone falls within a region for which a contig has already been assembled. Contig growth is also limited because small overlaps typically go undetected and some portions of the region being mapped may not have survived the cloning process. In fact, contigs assembled from cosmid clones typically stop growing after reaching lengths of 100 kbp.

Question: How do we order disconnected contigs along the chromosome and how do we check their accuracy?

Answer: Many types of lower-resolution maps can be used to position the contigs along a chromosome and to check that all the clones in a contig come from approximately the same region of the genome.

Example: The contigs constructed for yeast chromosomes, which had an average length of 100 kbp, were ordered relative to a high-density genetic-linkage map containing 400 markers spaced at an average physical distance of 30,000 base pairs. To check the integrity of each contig, the clones that form it were hybridized to very

Complete High-Resolution Restriction Map of Yeast Chromosome I



The high-resolution restriction map for yeast chromosome I was derived from a completed contig map of the chromosome. The Xs mark the beginning of the subtelomeric regions which are known to lie a few thousand base pairs away from the telomeres (ends) of the chromosome. Restriction sites for the thirteen-base cutter *Sfi*I and the eight-base cutter *Not*I and markers on the linkage map of chromosome I are localized to particular restriction fragments on the high-resolution restriction map. (Courtesy of Maynard Olson, Washington University)

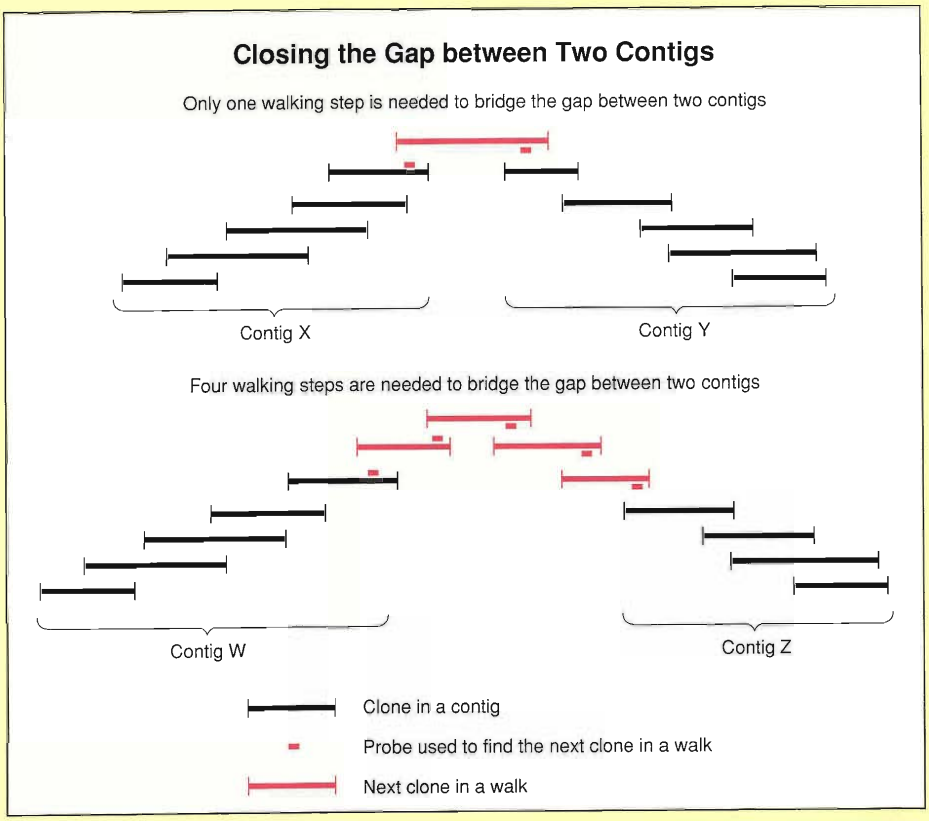
long (over 100,000 base pairs) restriction fragments of DNA that had been separated by pulsed-field gel electrophoresis. If the clones assigned to a contig do in fact come from a single region of the genome, it is likely that all of them will hybridize to a single large fragment on the gel.

The figure shows the high-resolution restriction map deduced from the completed contig map of yeast chromosome I. Also shown is the alignment of the restriction map with two other maps: (1) the genetic-linkage map and (2) a long-range restriction map showing the distances between the eight-base restriction sites of the enzyme *Not*I and the thirteen-base restriction site of *Sfi*I. (The latter map was constructed using pulsed-field gel electrophoresis.) Markers on the genetic-linkage map and restriction sites on the long-range restriction map have been localized to particular restriction fragments on the contig map. Those correspondences are indicated by dotted lines.

The contigs being assembled for human chromosomes are being checked by a variety of techniques including in-situ hybridization and hybridization to the DNA from hybrid cells containing increasingly longer portions of the chromosome being mapped (see "The Mapping of Chromosome 16").

Question: After the contigs are ordered and checked for accuracy, how do we fill in the gaps between the contigs?

Answer: As mentioned earlier, the fingerprinting of randomly selected clones is not an efficient way to fill in the gaps between contigs after the existing contigs cover a large fraction of the region being mapped. Instead it is time to employ a directed strategy. One directed strategy involves identifying unique regions within the clones at the ends of a contig and using those regions as probes to pick out other clones that will extend the contig. If the contigs cover a very large fraction (95 percent) of the region being mapped, a single probe from the end of a clone may identify a new clone that spans the distance between two existing contigs and thus merges them into one. If not, then one must continue stepwise by creating an end probe from each added clone and screening the library of clones to find the next clone that extends the contig a bit farther. This procedure is called walking, and it is extremely time-consuming. Nevertheless, it has been used successfully to complete physical maps of the *E. coli* and yeast genomes. Those genomes are relatively small (containing 5 million base pairs and 13 million base pairs, respectively), and the gaps between contigs were small before walking was attempted.



Example: The figure illustrates the merging of two contigs by either a single clone or several walking steps.

CAVEAT: A physical map is a very difficult puzzle to complete. As mentioned in the round table (see pages 108–109 in "Mapping the Genome"), the generic clone-to-fingerprint-to-contig cycle, which is amenable to automation and improved data-analysis algorithms, is only a small fraction of the work. The rest of the work required to close gaps between contigs and to track down inconsistencies such as the branching of one contig into two or more contigs involves many standard molecular-biology procedures, which, in the case of the human genome, must be carried out on an unprecedented scale. It is estimated that the completion of the yeast map took about 20 person-years of work, and the mapping of *each* human chromosome will take about 100 person-years. Further, mapping of human chromosomes presents some new challenges.

- An average human chromosome is ten times the size of the yeast genome, and the increased size calls for more efficient mapping strategies, such as working with larger clones.
- Unlike the genomes of yeast and *E. coli*, human DNA contains repetitive elements that require a new fingerprinting strategy to avoid inferring overlaps between clones containing long repetitive stretches of DNA near their ends.
- Experience has shown that regions containing repetitive sequences are often lost in the cloning process. Consequently, parts of the puzzle of each human chromosome may be missing, in which case completion of the map will require specialized techniques.

These challenges are being met in a variety of ways including the use of YAC cloning vectors, which accept DNA inserts eight to ten times larger than the inserts accepted by cosmids, and the use of STS markers, which, unlike restriction-fragment fingerprints, identify unique landmarks on the map and therefore eliminate the need for complicated probabilistic analyses to infer overlap between two clones. ■

STS Markers—a common language for mapping

David Botstein: One new thing has come out of the Genome Project itself and the way it's organized, to wit, the STS idea, which was thought up by Maynard in response to a strategic problem. The strategic problem was how to connect all the physical maps together, how to get a universal language. There was no way to make sure that everybody uses the same name for the same region of the genome because everybody was using different methods.

Then Maynard came up with this brilliant idea of STSs, and what we did very well as a community was to get everybody to understand that here was a universal language that would benefit everyone. In absolutely record time a new idea was adopted by a whole group of people without any mandate from anybody. STSs are an enormous technical advance because the DNA is used to label itself. At least fifty labs are now using STSs.

Bob Moyzis: I've been quoted as saying that the STS idea was a conceptual breakthrough, and I believe that's true. Maynard, perhaps you would explain where the idea came from.

Maynard Olson: I saw large-scale physical mapping as a kind of Tower of Babel. People were subdividing the problem by chromosome and by chromosome region, and I saw us ending up with a bunch of contig maps expressed in completely incompatible languages. That is, each group was building a map from a different clone collection and was using its own method for detecting clone overlaps. Consequently, we would have no convenient way to compare or crosscheck the maps. Eventually, the maps would have to be done again by whatever method proved to be the most generic.

The STS idea was to annotate each contig map with a series of unique landmarks. Each landmark, each STS, is just a short stretch of DNA—between 100 and 200 base pairs long—whose base sequence is found to be unique. Since the landmark is specified by a unique sequence of base pairs, it is called a sequence-tagged site, or an STS, and it can be unequivocally recognized and at the same time amplified by using the polymerase chain reaction [PCR].

The five-year plan includes the goal of generating a series of STSs for each chromosome separated by about 100,000 base pairs, but a series of more closely spaced STSs could eventually merge into the complete sequence.

I see the language of STSs playing a role similar to that of the ASCII code in the computer field. Once an early war between alternate standards was over, the ASCII code was adopted as the standard binary representation of a standard set of characters that are available on a standard typewriter keyboard.

The ASCII character set is woefully inadequate for today's word-processing needs because its roots are in a world view oriented toward typing. So when I try to read someone else's document on my computer, I may see a superscript where they had intended to have a Greek letter. Nonetheless, the fact that there was an ASCII code made a monumental difference to our ability to share information.

Similarly, the STS language enables us to share mapping information. In the late 1980s, when the feasibility of the Genome Project was being evaluated, the NRC committee struggled over the issue of generating incompatible maps, and we discussed the desirability of a common language for mapping. But at that time we had no experimentally practical technique to establish such a language. When the PCR was developed, it gave us an experimentally practical method for recognizing unique landmarks on the genome, and the sequences of those landmarks are the universal mapping language.

The crucial feature of STSs is that they have unique sequences. In other words, if we determine that two clones, one from each of two different clone collections, contain the same STS, we have no doubt that they come from the same region of the genome. Although the PCR is now the cheapest reliable method to recognize unique landmarks, one can imagine finding better recognition methods. However, I have no doubt that unique landmarks for the genome are going to remain short stretches of DNA with unique sequences, which is the essence of the STS concept.

The yeast map was constructed without STSs, and so the de facto landmarks on that map are the restriction sites at the ends of the particular clones that we used in its construction. We also made a restriction map specifying the distances between restriction sites. But restriction sites are repeated over and over in the genome, so they fail the most rudimentary test of an informative landmark. For example, if I give you a sample of yeast DNA and ask you whether it contains *Eco*RI site number 2708, you can't tell me whether it does or not. The restriction map does provide a back-up that would help to order a

new set of clones, but a great deal of the original work would have to be done over to order a new set of clones.

In fact, almost all applications of the current yeast map are completely dependent on the clone collection used to make the map. Those clones are stored in a repository and have also been transferred to filters and sent out to laboratories where work on yeast is being done.

People use radioactively labeled DNA probes to pick out the clone on the filter that binds to the probe by complementary base pairing. Then they look at the database describing the map to see what part of the genome that clone comes from. That procedure works, but, say, ten years from now when we have the complete sequence of the yeast genome, our yeast clones will have become irrelevant.

It's going to be some decades before the human genome is sequenced completely. In this transitional period it would be a serious error to have the intermediate utility of the physical-mapping effort completely dependent on scores of clone collections, each created with one or the other of the different cloning systems now being used. The average lifetime of a cloning system is five years. By then we've generally found a better system. Not only would we have to store all those clones in vast repositories, but we'd have to create the clone collections over and over again because clones don't last forever. Every time you propagate a clone, you run the risk of losing some of the DNA insert and sometimes even gaining some DNA. So a map has to be based on something besides clones.

Norton Zinder: Biological entities don't last. Every time you handle them, it's trouble.

David Botstein: That point is worth amplifying. The original NRC report includes a whole set of references to a central storage place for DNA clones, what we used to call the Sears Roebuck of molecular biology. Those of us who are practical-minded were concerned that the Sears Roebuck could well have cost more than the research. It might have been our supercollider. But it became clear very quickly that if an STS for some region of the genome has been identified, you can use the STS to pick out, from your own collection of clones, the clone containing that region. So the Sears Roebuck had become not only unnecessary but also undesirable.

I saw large-scale physical mapping as a kind of Tower of Babel. People were subdividing the problem by chromosome . . . and I saw us ending up with a bunch of contig maps expressed in completely incompatible languages.

Bob Moyzis: The idea that you can use an STS to pick out a clone of interest has already been tested experimentally in a number of labs. The base sequence of the STS can be transmitted electronically from one computer to another—no shipment of any material is involved. Along with the base sequence comes a protocol for a specific PCR that allows you to determine whether or not that STS is present in any given DNA sample. In particular, the PCR can be used to screen a library of anonymous DNA clones to isolate the clone containing the STS. It is important to note that, although we all talk of *the* human genome, there are as many human genomes as there are humans. The region of the generic genome related to to some disease, for example, will have to be isolated from the DNA of many unaffected and disease-affected individuals in order to determine the DNA changes associated with the disease. STSs will be invaluable for that job.

Norton Zinder: The STS idea has another consequence: Anyone can play. Everyone thought you had to have a big lab to contribute to the Genome Project. But now anyone who provides an STS for a human DNA fragment with a reasonably well-defined location is contributing to the goals of the Genome Project.

Maynard Olson: When some of the large genome centers start producing long-range continuous maps, and by that I mean contigs that span millions of base pairs and that are separated by relatively small gaps, we're going to see a tremendous amount of detailed mapping in the smaller labs. People will hone in on particular regions of interest and add all kinds of annotations to the maps, not only new STSs, but also sites of translocations, mutations, genes, regulatory regions, and so forth.

We see that happening with the yeast map. People all over the place are annotating it in much the same way that road maps are annotated with information about the locations of parks, public buildings, and historic sites. Adding details to a map is easy compared with the initial construction of a map, but those details greatly increase the value of the map. In fact, to a large

degree, they create the value of the map. Relatively large efforts are under way to create the initial contig maps for human chromosomes, but the little players will also get to contribute by straightening out the regions of the physical maps that they know a lot about. We've got to get some of the long-range physical maps out there, and then the community will do a superb job of annotating them.

When I suggested we could convert physical maps into pure information that could be stored in a computer, I didn't mean to imply that STSs do the whole job and that people outside the Genome Project would have to make their own contigs from earth, air, fire, and water. We'll still be sharing clones and other information about the contig maps. The purpose of STSs is to provide some bedrock to build on, some unique landmarks, that let you know where you are on the genome.

Bob Moyzis: The Genome Project's five-year goals for physical mapping are, in fact, to create a map for each human chromosome made up of contigs that are between 1 million and 2 million base pairs in length and that cover 95 percent of the chromosome, and second, to generate STSs spaced at intervals of 100,000 base pairs along each chromosome.

So we have to generate about 30,000 STSs. Some investigators initially thought that STS generation was an additional burden. However, now most laboratories consider STS generation to be a trivial part of physical mapping—a minor fraction of their total costs.

David Botstein: Generating an STS involves preparing a short DNA segment for sequencing, determining its base sequence, picking out unique primer sequences for the PCR, and then testing the PCR for its ability to recognize and amplify that unique stretch of DNA.

We don't have a good estimate of how much it costs to generate each STS because people are generating STSs under different conditions and everybody's overhead is different. It's probably simpler to talk about the time required. I would say that if a person starts with a piece of cloned DNA, he can generate an STS from that clone in about two weeks.

The STS idea has another consequence: Anyone can play . . . anyone who provides an STS for a human DNA fragment with a reasonably well-defined location is contributing to the goals of the Genome Project.

Bob Moyzis: Yes, David, but one person can process many STSs in parallel. It's been our experience at Los Alamos that one person can generate approximately a hundred STSs per year. Therefore, generating each STS costs approximately a thousand dollars. And as I said, that is a small fraction of the total cost of physical mapping.

STSs are a means of annotating a contig map that's already constructed, but now that everybody's begun to accept the language of STSs, those markers are also being used as a primary means for building contigs, for detecting whether two clones overlap. And unlike restriction-fragment fingerprints, which give only the probability of overlap, the presence of an STS in two clones is a guarantee of overlap.

So mapping can be carried out by first identifying a bunch of STSs and then finding pairs of clones containing the same STS. Clones that share an STS must overlap and thus belong in the same contig. This approach, called STS-content mapping, was pioneered at Washington University. It has become the approach used by most genome centers to construct the contig maps and to distribute the information in those maps.

David Botstein: That's been a major technological change in one year. We were looking at a lot of restriction-fragment fingerprint experiments a year ago, and now we're looking at a lot of STS-content experiments that are doing exactly the same thing, namely aligning one piece of DNA with another piece by detecting that the two overlap. There's no question that the STS-content paradigm is now the standard for physical mapping.

Bob Moyzis: That's the most efficient method for constructing a low-resolution map consisting of YAC-sized clones with unique landmarks spaced at intervals of 100,000 base pairs. But the most efficient method for creating a higher-resolution map with landmarks every few thousand base pairs is to fingerprint cosmid clones and create cosmid contigs covering the region within each YAC. We are geared to do that at Los Alamos and have found that it takes approximately two weeks to convert a YAC into a cosmid contig. Generating STSs at every few thousand base pairs would be much more work, at least by current methods.

STSs and Genetic Linkage Maps

Norton Zinder: Genetic-linkage maps can also be expressed in the language of STSs. All we have to do is generate an STS for each polymorphic DNA marker by sequencing each marker and developing a PCR to amplify a unique sequence within the marker.

Our five-year goal for the genetic-linkage maps is to find markers spaced evenly along the genome at genetic distances of 2 to 5 centimorgans, which translates into physical distances of 2 to 5 million base pairs. So we need about 600 polymorphic STSs—if they're equally spaced—to give us a pretty good genetic map of the genome, and we will need about ten times that number if we develop those markers from randomly chosen clones.

Nancy Wexler: People looking for disease genes probably haven't stopped to make an STS for each polymorphic DNA marker they are working with. We have discussed offering people some incentive either to do that themselves or to send their polymorphic DNA markers to some central place. In any case, the number of STSs is going to increase. The beauty of STSs is that they save real estate because you don't have to store clones.

Bob Moyzis: I expect that many of the polymorphic STSs will be generated around particular disease loci because PCRs can then be used to isolate the DNA of those variable sites directly from many patients. So we're going to have many more STSs than the number that is specified in the five-year plan. They may not, however, be generated with the desired spacing.

David Botstein: We need to remind people that there's no point in making a physical map if you don't have a high-density genetic-linkage map, that is, one on which the polymorphic markers are closely spaced.

Nancy Wexler: And only through the linkage map can we infer that a gene for a particular inherited trait is located near a particular marker. The folks working on genetic diseases are waiting avidly for the linkage maps.

Genetic-linkage maps can also be expressed in the language of STSs . . . We need about 600 polymorphic STSs—if they're equally spaced—to give us a pretty good genetic map of the genome, and we'll need ten times that number if we develop those markers from randomly chosen clones.

Norton Zinder: At the moment most groups working on disease genes are retaining only those markers that turn out to be closely linked to the gene of interest, and they're discarding other markers that they come across. That's rather inefficient because the discarded marker might be relevant to genes in another region.

Nancy Wexler: People working on the same region often come up with different linkage maps, but they don't necessarily get around to resolving the differences. The managers of the Genome Project say the goal is to expedite getting the most accurate linkage map, and your funding is dependent on your sitting down with a committee and figuring out what experiments to do to resolve the discrepancies and connect the maps. That incentive seems to be working quite well.

Lee Hood: At Caltech we are developing automated techniques for genetic mapping. Present methods for identifying polymorphic DNA markers generally require gel electrophoresis and are therefore hard to automate. We are developing and automating an assay, the oligonucleotide ligase assay [OLA], which can readily identify known polymorphisms, in particular, those involving single-base changes.

The assay employs two DNA probes, about 20 bases long, that are complementary to adjacent regions in the genome. The polymorphism detected by the array includes the base at the 3′ end of the 5′ probe—the base directly adjacent to the 3′ probe. The 5′ probe has biotin attached to its 5′ end; and the 3′ probe has a reporter group at its 3′ end. When the two probes are hybridized to the target DNA, DNA ligase will covalently join them if and only if there is perfect molecular complementariness between the probes and the target sequence. The sequences containing biotin are then pulled from the reaction mixture and assayed for the presence of the 3′ reporter group.

If there is an exact match between the probes and the target sequence, then the 3′ reporter group will be present on the biotin-labeled sequences that are pulled out of the mixture. If there is no match, then only the 5′ probe will

be pulled out from the mixture. Hence the assay is a simple plus or minus assay for the presence of a particular form [allele] of a polymorphism. A second 5′ probe can be synthesized complementary to the second allele of the polymorphism—and the same DNA can be assayed again. Thus we can determine whether an individual is homozygous or heterozygous for that polymorphism.

We are in the process of automating this entire procedure with a robotic work station. A single person can analyze 1200 assays in a day. This reaction can be carried out in the individual wells of a 96-well microtiterplate. We first amplify the target sequence in each well using PCR and then use the ligation assay. We're developing techniques for rapidly determining polymorphisms with two alleles so that OLA can be used to map entire chromosomes.

We're also working with Los Alamos to generate markers for chromosome 14 using a chromosome-specific library of clones. We're randomly sequencing cloned fragments from the library, picking out those regions from the DNA of six individuals and sequencing those regions again to identify frequent polymorphisms.

We have found that three or four polymorphisms often fall within a thousand base pairs and that these closely spaced markers are in partial linkage equilibrium, so that they provide highly informative markers for linkage analysis. In a relatively short period of time, we've generated seven such markers.

Now that a technique for identifying those markers in the DNA from any individual, namely, OLA, is semi-automated, if you want to use those markers to identify the relative position of a particular trait on chromosome 14, you can readily do it.

Bob Moyzis: It's important to point out that if the genetic information obtained by this project is to be widely utilized, then automated techniques similar to those Lee just described must be developed.

Screening the whole genome for markers that are linked to a particular disease is still a very painful process for most laboratories, in part because those markers are not collected in any single place. That's why the Genome Project has decided to produce a kit of 150 reference markers spaced evenly over the genome at distances of about 20 million bases . . . the Project can create an appropriate infrastructure and deliver the goods to the scientific community.

David Cox: That's right, but in the meantime, screening the whole genome for markers that are linked to a particular disease is still a very painful process for most laboratories, in part because those markers are not collected in any single place.

That's why the Genome Project has decided to produce a kit of 150 reference markers spaced evenly over the genome at distances of about 20 million bases. That effort involves identifying and collecting existing markers and supporting various individuals to search for probes in regions where no probes yet exist.

Those regions will be targeted by radiation-hybrid mapping or microdissection of chromosomes. Once the markers are collected, they will be put together in a package that will be sold at a reasonable price. Ray White and Helen Donis-Keller have each constructed a genetic-linkage map for the human genome. They say that they're happy to let other people have their markers. They just don't have the time and the money to distribute them.

Other available markers come with strings attached and rightly so because private companies have put a lot of money into generating them. The labs that have done the most work on the genetic map are often criticized for not sharing, but in many instances, they just don't have the infrastructure that allows them to share efficiently. On the other hand, the Genome Project can create an appropriate infrastructure and deliver the goods to the scientific community. The reference list is an immediate goal that can be fulfilled.

We must remember that the Genome Project is a product-oriented endeavor, and those who are funded are expected to come through with the product. In normal research, you can't always predict exactly what you're going to find, but here we have very specific goals.

The Polymerase Chain Reaction and Sequence-tagged Sites

Norman A. Doggett

Polymerase Chain Reaction

The polymerase chain reaction (PCR) is an in vitro method for selectively amplifying, or synthesizing millions of copies of, a short region of a DNA molecule. The reaction is carried out enzymatically in a test tube and has been successfully applied to regions as small as 100 base pairs and as large as 6000 base pairs. In contrast, DNA cloning is a nonselective in vivo method for replicating DNA fragments within bacterial or yeast cells. Cloned fragments range in length from several hundred to a million base pairs. (See "DNA Libraries" for further discussion of DNA cloning.)

PCR is particularly important to the Human Genome Project as a tool for identifying unique landmarks on the physical maps of chromosomes. The PCR can be used to detect the presence of a particular DNA segment in a much larger DNA sample and to synthesize many copies of that segment for further use as a probe or as the starting material for DNA sequencing.

Figure 1 illustrates the polymerase chain reaction. The reaction mixture contains:

- A DNA sample containing the target sequence.
- Two single-stranded DNA primers (short sequences about 20 nucleotides long) that anneal, or bind by complementary base pairing, to opposite strands of DNA at sites at either end of the target sequence. Such short DNA sequences are called oligonucleotides and can be synthesized in a commercially available instrument.
- A heat-stable DNA polymerase, an enzyme that catalyzes the synthesis of a DNA strand complementary to the target sequence and can withstand high temperatures.
- Free deoxyribonucleoside triphosphates (dATP, dGTP, dCTP, and dTTP), precursors of the four different nucleotides that will extend the primer strands.
- A reaction buffer to facilitate primer annealing and optimize enzymatic function.

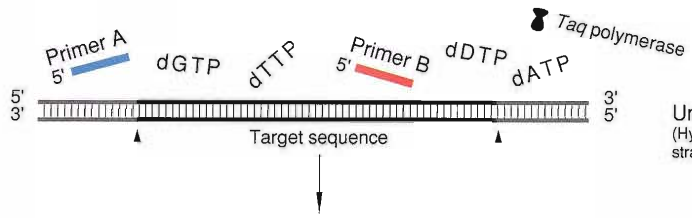
The polymerase chain reaction proceeds by repeated cycling of three temperatures:

- Phase 1: Heating to 95°C to denature the double-stranded DNA, that is, to break the hydrogen bonds holding the two complementary strands together. The resulting single strands serve as templates for DNA synthesis.
- Phase 2: Cooling to a temperature between 55°C and 65°C to allow each of the primers to anneal (or hybridize) to its complementary sequence at the 3′ end of one of the template strands.
- Phase 3: Heating to 72°C to facilitate optimal synthesis, or extension of the primer strand by the action of the DNA polymerase. The polymerase attaches at the 3′ end of the primer and follows along in the 3′-to-5′ direction of the template strand catalyzing the addition of nucleotides to the primer strand until it either falls off or reaches the end of the template strand (see "DNA Replication" in "Understanding Inheritance").

The figure shows the materials in the reaction mixture and the first three cycles of the reaction. The DNA synthesized in each cycle serves as a template in the next. Note that an exact duplicate of each strand of the target sequence is first created during cycle 2. Each subsequent cycle doubles the number of those strands so that after n cycles the reaction will contain approximately 2^n copies of each strand of the

Figure 1. The Polymerase Chain Reaction

Reaction mixture includes DNA sample: two single-stranded primers, each with a 20-base sequence identical to the 5' end of one strand of the target sequence; heat stable *Taq* polymerase; and deoxyribonucleocide triphosphates (dNTPs).

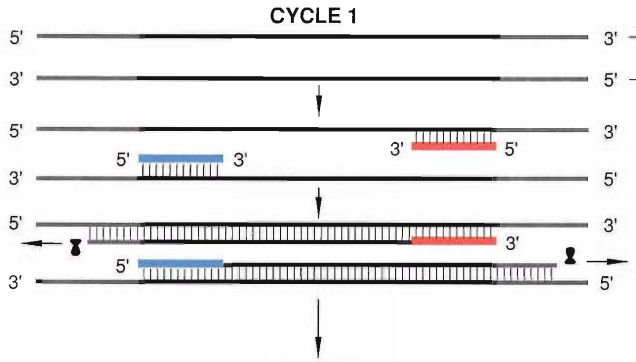


Unamplified DNA
(Hydrogen bonds between the two DNA strands are shown as vertical lines)

Phase 1
Denature unamplified DNA at 95°C to form single-stranded templates.

Phase 2
Anneal primers to template at about 60°C.

Phase 3
Synthesize new strands at 72°C.



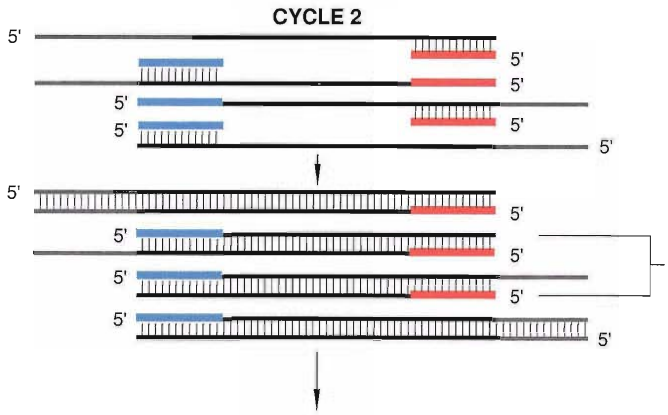
Template strands

Each primer binds to its complementary sequence at the 3' end of a strand of the target sequence.

Each new strand is synthesized in the 5'-to-3' direction.

Phases 1 and 2
Denature products of Cycle 1 and anneal primers to template strands.

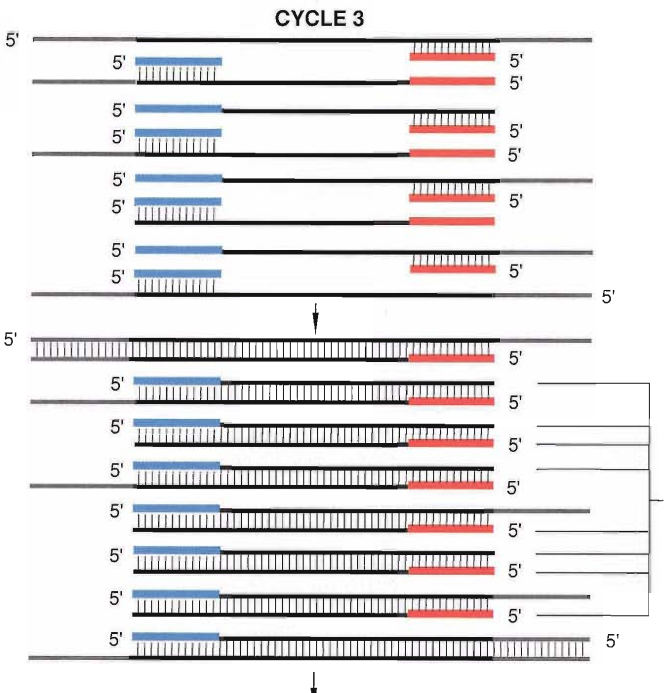
Phase 3
Synthesize new strands.



First two template strands that span the target sequence only

Phases 1 and 2
Denature products of Cycle 2 and anneal primers to template strands.

Phase 3
Synthesize new strands.



Eight template strands that span the target sequence only

Continue for 20 to 30 cycles to produce over 10^6 copies of target sequence.

target sequence. Typically the chain reaction is continued for 20 to 30 cycles in microprocessor-controlled temperature-cycling devices to create between roughly 1 million and 1 billion copies of the target sequence.

Taq polymerase, a heat-stable polymerase isolated from the bacterium *Thermus aquaticus* found in hot springs, is used in the reaction. The annealing temperature for the second phase of each cycle is chosen to be approximately 5°C below the temperature at which the primers no longer anneal to the target sequence. That so-called melting temperature varies depending upon the primer sequence. In particular because G-C base pairs (which have three hydrogen bonds) remain stable at higher temperatures than A-T base pairs (which have only two hydrogen bonds), primers containing mostly Gs and Cs have a higher melting temperature than those containing mostly As and Ts. The annealing temperature must be chosen carefully because if the temperature is too low, the primers will bind to sites whose sequence is not exactly complementary to the primer sequence resulting in the amplification of sequences other than and in addition to the target sequence. If the temperature is too high, the primers will not bind to the template strands and the reaction will fail.

Typically the initial DNA sample contains from 3,300 to 333,000 copies of the human genome (or 10 nanograms to 1 microgram of total genomic DNA). However, when working properly, the PCR will selectively amplify a unique target sequence contained in a single copy of the genome (6 picograms of DNA) isolated from a single cell. To evaluate the specificity of the reaction, that is, whether or not the reaction amplified a single target region, the reaction products are separated on a gel using electrophoresis. If a single region has been amplified, the gel will contain a single intense band containing the synthesized copies of the target sequence. The location of the band on the gel indicates the length of the amplified region. If more than one intense band appears on the gel, then more than one region of the genome was amplified by the reaction and the sequence of the primers appear more than once in the genome.

Sequence-tagged Sites

A sequence-tagged site (STS) is a short region along the genome (200 to 300 bases long) whose exact sequence is found nowhere else in the genome. The uniqueness of the sequence is established by demonstrating that it can be uniquely amplified by the PCR. The DNA sequence of an STS may contain repetitive elements, sequences that appear elsewhere in the genome, but as long as the sequences at both ends of the site are unique, we can synthesize unique DNA primers complementary to those ends, amplify the region using the PCR, and demonstrate the specificity of the reaction by gel electrophoresis of the amplified product (Figure 2).

Operationally, a sequence-tagged site is defined by the PCR used to perform the selective amplification of that site. The PCR is specified by the pair of DNA primers that bind to the ends of the site and the reaction conditions under which the PCR will amplify that particular site and no other in the genome.

STSs are useful because they define unique, detectable landmarks on the physical map of the human genome. One of the goals of the Human Genome Project is to find STS markers spaced roughly every 100,000 bases apart along the contig map

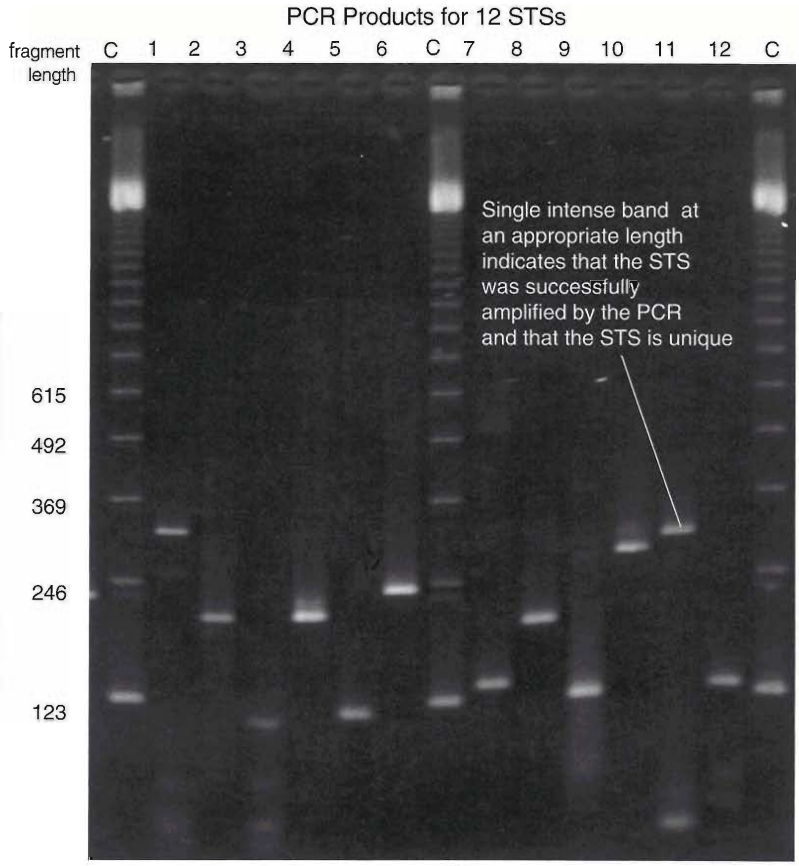
of each human chromosome (see "Physical Mapping—A One-dimensional Jigsaw Puzzle" for a description of contig maps). The information defining each site will be stored in a computer database such as GenBank. That stored information will include the PCR primers, reaction conditions, and product sizes as well as the DNA sequence of the site. Anyone who wishes to make copies of the marker would simply look up the STS in the database, synthesize the specified primers, and run the PCR under the specified conditions to amplify the STS from genomic DNA. As described below, copies of the STS can be used to screen a library of uncharacterized clones and identify a clone containing the marker. Therefore, a database of such landmarks will eliminate the need to store and distribute a permanent set of DNA clones or probes for the physical maps.

Figure 3 outlines the procedure for finding an STS marker. One begins by sequencing a 200- to 400-base region of a cloned DNA fragment. The rough sequence can be obtained from a single run of a DNA sequencing gel (see "DNA Sequencing"). The sequence is then examined to find two twenty-base regions separated by 100 to 300 base pairs that might serve as unique primers for a PCR (see Figure 4). The primers are synthesized and then the PCR reaction is run on genomic DNA to see whether the reaction results in the selective amplification of the targeted region. If it does, then the amplified region becomes an STS. In our work at Los Alamos, we found that about half of the sequences we obtained from randomly selected clones yielded an STS.

STS Markers for Physical Mapping

STSs are being used to find pairs of overlapping clones for the construction of contig maps of human chromosomes. Since each STS is a unique site on the genome, two clones containing the same STS must overlap and the overlapping region must include the STS.

Figure 2. STSs from Chromosome 16



Electrophoretic Gel

To check that a sequence-tagged site (STS) is a unique sequence on the genome, the polymerase chain reaction (PCR) defining that site is carried out on total genomic DNA and the products of the reaction are separated on a gel by electrophoresis. If the reaction amplifies that site and no other, all reaction products will have the same length (known from the sequence of that site) and will appear together as a single intense band on the gel.

At Los Alamos, twelve dIfferent STS markers are amplified in parallel by the PCR, and the products are separated on twelve separate lanes of a gel. The presence of only one intense band in each numbered lane of the gel shown above indicates that the STS is indeed a unique site. Fainter bands near the bottom of a lane are residual primers remaining after the PCR. The sizes of the amplified products are measured relative to a ladder of standard fragments (with known lengths that are multiples of 123 base pairs) that have been separated by length in the gel lanes marked C.

Before overlap can be detected, clones containing the same STS must be identified from among a collection of clones in a DNA library. If the individual cloned fragments have been permanently arrayed on nitrocellulose or nylon membranes,

Figure 3. Steps in Developing an STS Marker

Either create a chromosome-specific library of M13 clones, or pick a clone from the end of a cosmid contig, digest the cosmid clone with a restriction enzyme, and clone the restriction fragments in M13 cloning vectors.

↓

Sequence 200 to 400 base pairs of DNA from an M13 clone. The rough sequence determined from a single run on a DNA sequencing machine is sufficient for identifying an STS. (By "rough" we mean an average error rate of 1 in 100 bases.)

↓

Compare the sequence to all known repeated sequences using computer algorithms to help identify regions likely to be unique.

↓

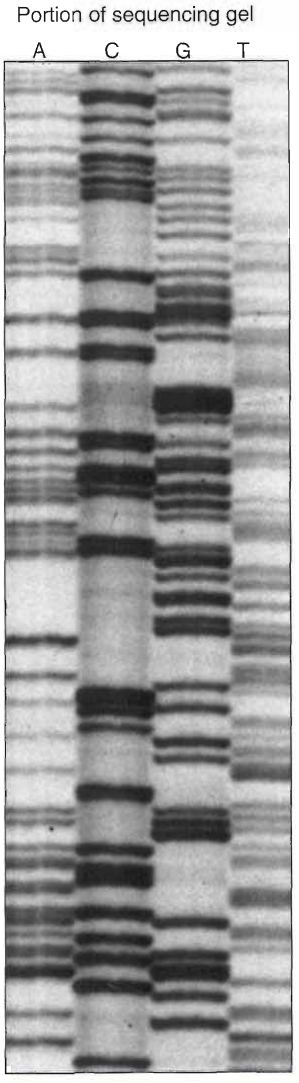
Select two primer sequences from the unique regions that are separated by 100 to 300 base pairs. Gs and Cs should comprise 45 to 55 percent of the bases in each primer sequence, and the melting temperatures of the two primers should differ by less than 5° C (see example in Figure 4).

↓

Synthesize the primers and use them to run the PCR on genomic DNA isolated from human cells. Analyze the amplification products by agarose gel electrophoresis to evaluate the specificity of the reaction.

↓

A functional STS marker will amplify a single target region of the genome and produce a single band on an electrophoretic gel at a position corresponding to the size of the target region.



then clones containing a particular STS may be identified by hybridization to copies of an STS marker. First, copies of the STS are generated from genomic DNA by the PCR. The amplified copies are labeled with radioactive ^{32}P, denatured, and then applied to the membranes containing the arrayed collection of cloned fragments. The labeled markers will hybridize only to those clones containing DNA sequences complementary to those of the markers. Clones that are positive for the STS are imaged as dark spots on x-ray films that have been exposed to the membranes containing those clones.

A more rapid screening method involves dividing a library of clones into pools and using PCR to interrogate each pool for the presence of the STS. In the PCR-based screening method, primers are synthesized for each STS, and many pools are screened in parallel. If a particular pool of cloned fragments supports PCR amplification of the STS target sequence, then at least one particular clone in the pool must contain the target sequence. Using a clever pooling scheme described below, the identification of which pools support amplification will result in the identification of the particular clone or clones containing the STS.

STS Markers for the Chromosome-16 Physical Map

In line with the five-year goals of the Human Genome Project, the Los Alamos effort to construct a physical map of chromosome 16 includes developing STS markers spaced, on average, at 100,000-base-pair intervals along the chromosome. At present about 60 percent of chromosome 16 is covered by contigs made up of cosmid clones. On average each cosmid contig spans a distance of 100,000 base pairs. We are developing STSs by sequencing regions from the clones that lie at either end of each contig. Thus far a total of 325 sequences have been obtained from such clones and about 100 of these have been developed into STSs. The STS markers will be stored in GenBank so that anyone who wants to regenerate the markers and use them to identify clones containing those markers may do so.

The STS markers from the end clones of our cosmid contigs are serving several purposes. First, they are being used to screen a library of YAC clones for clones that may overlap two different cosmid contigs and therefore close the gap between them.

Our library of 550 YACs is specific for chromosome 16. That is, the YACs contain DNA inserts from human chromosome 16 only. Since those inserts have an average size of 215 kb, the total YAC library represents a one-time coverage of the DNA in chromosome 16. The construction of such chromosome-specific YAC libraries is an important breakthrough for physical mapping and is described in "Libraries from Flow-sorted Chromosomes."

Figure 4. Example of an STS

Rough Sequence—347 Bases (lower case letters indicate uncertainty in the base call)

```
                                                    ┌ Primer A
5   5'-GAATTCCTGA  CCTCAGGTGA  TCTGCCCGCC  TCGGCCTCCC  AAAGTGCTGG
51     GATTTACAGG  CATGAGGCAC  CACACCTGGC  CAGTTGCTTA  GCTCTCTAAG
101    TCTTATTTGC  TTTACTTACA  AAATGGAGAT  ACAACCTTAT  AGAACATTCG
151    ACATATACTA  GGTTTCCATG  AACAGCAGCC  AGATCTCAAC  TATATAGGGA
201    CCAGTGAGAA  ACCAATGTCA  GGTAGCTGAT  GATGGGCAAa  GGgATGGGgA
251    CTGATATGCC  cNNNNNGACG  ATTCGAGTGA  CAAGCTACTA  TGTACCTCAG
301    CTTTtCATCT  tGATCTTCAC  CACCCATGGg  TAGGTGTCAC  TGAAaTT-3'
                3'-CTAGAAGTG  GTGGGTACCC  AT-5' ──── Primer B
```

		Melting Temperature
Primer A	5'-GTT TCC ATG AAC AGC AGI CAG-3'	69.4°C
Primer B	5'-TAC CCA TGG GTG GTG AAG ATC-3'	68.7°C

The STS developed from the rough sequence shown above is 171 bases long. It starts at base 162 and runs through base 332. Primer A is 21 bases long and lies on the sequenced strand. Primer B is also 21 bases long and is complementary to the shaded sequence toward the 3' end of the sequenced strand. Note that the melting temperatures of the two primers are almost equal. A computer algorithm was used to pick out the two primer sequences and to calculate their melting temperatures.

We have partitioned the YACs into pools and are using a PCR-based screening strategy to identify YACs containing each STS. Our pooling scheme, devised by David Torney in the theoretical biology group at Los Alamos, has the advantage of detecting false positive and false negative results from the PCR (see "YAC Library Pooling Scheme"). Once a YAC clone containing an STS is identified, a PCR technique (known as inter-ALU PCR) is used to generate a set of probes from that YAC. The probes are hybridized to our arrayed library of cosmid clones. If clones from two different contigs yield positive hybridization signals, then the YAC must bridge the gap between the two contigs. So far we have identified 30 YACs containing the STSs from end clones of cosmids. These YACs and seventy-five others have been hybridized to the cosmid clones resulting in the closure of sixty-five gaps in the contig map of chromosome 16..

The same STSs are being used to localize each of our cosmid contigs to an interval on chromosome 16, defined by a series of mouse/human somatic-cell hybrids containing various portions of chromosome 16. Collaborators David Callen and Grant Sutherland of Adelaide Children's Hospital in Southern Australia have collected a panel of 50 hybrid cells that divide chromosome 16 into 50 intervals with an average size of 1.7 million bases. Using a hybridization-based method and, more recently, our STSs and a PCR-based strategy, they have screened the DNA in each hybrid cell and thereby localized each of 70 contigs to one of the 50 intervals defined by the hybrid-cell panel. Those 70 contigs represent about 10 percent of chromosome 16.

STS Markers for Genetic-linkage Mapping

So far we have suggested that an STS yields the same product size from any human DNA sample. However, STSs can also be developed for unique regions along the genome that vary in length from one individual to another. The PCR that amplifies

Figure 5. Polymorphic STSs—Highly Informative Markers for Linkage Analyis

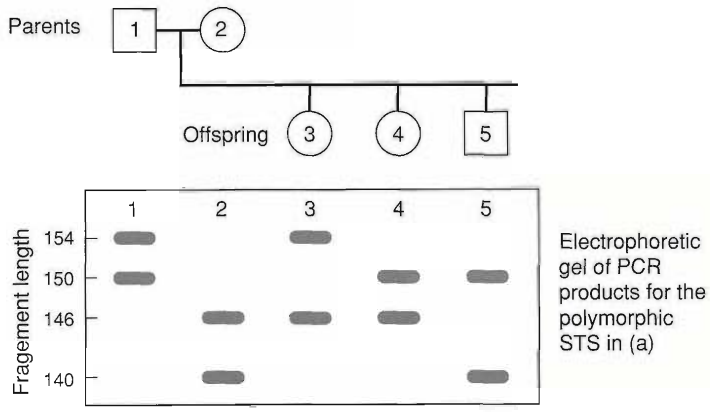
(a) A Polymorphic STS

5'- unique sequence GTGTGT . . . . . GT unique sequence -3'

$(GT)_n$

The number *n* of GT repeats varies among the population.

(b) Inheritance of the Polymorphic STS shown in (a)



A variable locus containing a short repeated sequence, such as the dinucleotide repeat $(GT)_n$, flanked by two unique sequences can be developed into an STS. An example is shown in (a). The size of the amplified product for that STS will vary depending on the value of *n* at that locus, and therefore the STS is polymorphic. Each individual carries two copies of the STS marker, one on each chromosome of a homologous pair, and each copy may have a different value of *n* and thus be a different allele of the polymorphic STS.

The inheritance of the polymorphic STS in a five-member family is illustrated schematically in (b). The electrophoretic gel shows the PCR products for the STS from each family member. The two alleles carried by the father are different from the two alleles carried by the mother. The children inherit one allele of the STS from each parent.

Because markers developed around such repeat sequences have many alleles, the likelihood that a given individual is heterozygous for such a marker is high. As explained in "Classical Linkage Analysis," at least one parent must be heterozygous for two different markers (or genes) in order to establish linkage between the two. Thus markers that have many alleles are likely to be *highly informative* for linkage analysis. (See "Informativeness and Polymorphic DNA Markers.") Polymorphic STSs will help to attain the five-year goal to construct a genetic-linkage map of highly informative DNA markers spaced at genetic distances of 2 to 5 centimorgans along each chromosome of the human genome. Moreover, these STSs are easily located on the physical map and thus provide a convenient means for aligning the linkage map with the physical map of a chromosome.

the variable region will yield different product sizes depending on which variations of the region are present in the genome of a given individual. An STS from a variable region is, by definition, a polymorphic DNA marker, which can be traced through families along with other DNA markers and located on genetic-linkage maps (see "Modern Linkage Mapping").

Figure 5(a) shows an example of a unique region that has variable lengths and can be developed into a polymorphic STS. At either end of the region is a unique sequence about 20 nucleotides long that can serve as a primer sequence for the PCR. Between those two sequences is a simple tandem repeat, $(GT)_n$ (or GT repeated in tandem *n* times). Such dinucleotide repeats are scattered throughout the human genome as are tri-, tetra-, and pentanucleotide repeats. Moreover, the number *n* of tandem repeats at a given locus along a chromosome is an inherited trait that tends to vary widely among the population. Thus each such variable locus has many different alleles (or forms), each one defined by the number *n* of tandem repeats between the unique sequences.

STSs are being developed for this abundant class of variable regions. Since the varying sizes of the PCR products from a polymorphic STS correspond to the alleles of that marker, PCR followed by gel electrophoresis of the amplified products is the method of detecting which alleles of the marker are carried by an individual [see Figure 5(b)].

Polymorphic STSs are particularly useful because they can serve as landmarks on both the physical map and the genetic-linkage map for each chromosome, and thus they provide points of alignment between the different distance scales on these two types of maps.

At Los Alamos we have identified the location of $(GT)_n$ repeats as part of our fingerprinting and mapping strategy (see "The Mapping of Chromosome 16"). We are now developing these regions into STSs for use in linkage mapping. ■

YAC Library Pooling Scheme for PCR-Based Screening

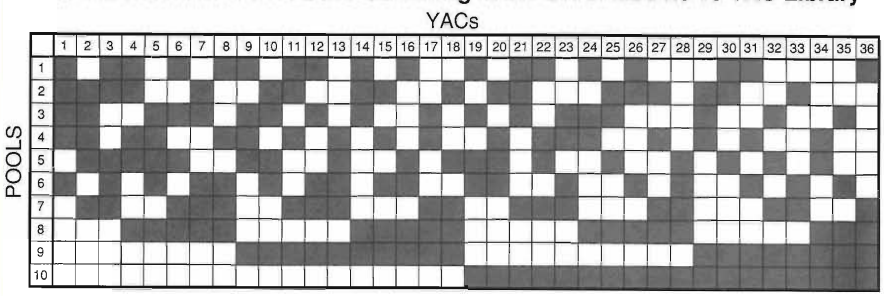
David J. Balding and David C. Torney

The PCR is a rapid method for screening a library of clones for the presence of clones containing an STS. Usually the library is divided into pools of clones, and the PCR is run on each pool. The problem we address here is to design efficient and robust pooling schemes for such PCR-based screening. Two questions are relevant: (1) Given an arbitrary unique sequence, how should one pool a library of clones to find rare positive clones (those containing this unique sequence), using a reasonable number of pools and a minimum number of pools queried per positive? (2) How can the design of the pooling scheme be robust to experimental errors (false positives, false negatives) when querying pools with PCR? Clearly, we want to do group testing in a way that gives correct results even in the presence of experimental errors.

In answer to these questions, we designed a pooling scheme called a J-detector, capable of indicating either which j clones are positive for $j \leq J$, or whether more than J clones are positive. The scheme works in the presence of K experimental errors provided any one clone in the J-detector occupies at least $K+1$ pools that are not among the pools jointly occupied by any set of J other clones. For example, if $J = 1$, and $K = 0$, we require that, among the pools containing clone_i, there is at least one pool that does not contain clone_j for all $i \neq j$. Thus we can distinguish one positive from two positives.

From information theory we know that the number of pools in a J-detector must be at least $J \log N$, where N is the number of clones in the library. We believe that t-designs (Beth et al., 1986) constitute optimal J-detectors, therefore we focused our efforts on improved methods for the construction of t-designs. A t-design has three parameters: v, the number of pools; k, the number of pools each clone occupies; and t, the maximum number of pools any two clones jointly occupy.

The 1-Detector for PCR-Based Screening of the Chromosome 16 YAC Library



The design for a 1-detector containing 36 YAC clones has ten pools. Each clone is in 5 pools and two clones occur jointly in no more than 3 pools.

The chromosome 16-specific YAC library developed at Los Alamos contained 550 clones with an average insert size of approximately 215 kb, representing approximately a one-fold coverage of this chromosome. We chose to divide the library into 16 partitions each containing 36 clones and construct a 1-detector with $K = 1$ for the clones in each partition. In other words, the pooling scheme allows us to detect (1) which single clone among the 36 is positive for an STS or (2) whether there is more than one positive clone in the partition, even in the presence of an erroneous PCR reaction.

Assuming our YAC library represents uniform one-fold coverage of chromosome 16, the probability that more than one positive will occur in any of the 16 1-detectors is approximately 0.01. These 1-detectors (shown in the figure) are given by the t-design with parameters $v = 10$, $k = 5$, and $t = 3$. Note that the five pools containing one clone and the five pools containing another clone have at most three pools in common as $t = 3$.

Suppose only one clone in a 1-detector is positive for a given STS. Then even if one pool containing the positive clone yielded a false negative and only four pools containing that positive clone yielded positive results, one could use parsimony to tentatively identify the positive clone ($K = 1$). If the 1-detector contained two positive clones, at least seven pools would yield positive results (in the absence of experimental errors), a result readily distinguished from the five positive pools expected for a single positive clone. In fact, 4/7 of the time, only seven pools would be positive and all but three clones would be identified as negative. Thus, even when more than one clone in the 1-detector is positive for a given STS, the screening identifies a large number of negative clones, which can be eliminated from further consideration.

To identify which of the sixteen 1-detectors to screen, one could implement two levels of a four-way branching tree like that of Green and Olson (1990). Then, a maximum of 20 PCR reactions are required to identify each positive clone. Our pooling scheme has been successfully used to identify 30 YACs each containing a different STS. In almost all cases, PCR screening for each STS yielded five positive pools in a 1-detector, and the clone thereby identified as positive was always confirmed in subsequent analysis.

We plan to take advantage of the larger t-designs in future experiments. For example, the design with $v = 12$, $k = 6$, and $t = 4$ will accommodate 132 clones in its 12 pools. We found that the Biomek robot can create these pools given a bit-string representation. ■

Sequencing cDNAs

Bob Moyzis: Since the Five-Year Plan was written, one of the new proposals that has surfaced is to sequence a large number of complementary DNAs, or cDNAs. Let's discuss the rationale behind this approach.

David Galas: Once it became clear that Maynard's idea about STSs was going to be a fruitful way to deal with physical mapping, the question arose: As long as we have to sequence short stretches of DNA to make STSs, why not choose those short stretches from cDNAs rather than from some random set of DNA fragments? cDNAs are interesting because they are copies of genes that actually get expressed as proteins in human cells.

We make cDNAs by isolating messenger RNAs, or mRNAs, from cells and using the enzyme reverse transcriptase to change those protein-synthesis templates back into the DNA message. But unlike the original DNA message, the cDNAs do not contain the noncoding regions, called introns, because RNA splicing has removed them [see "Gene Expression and cDNAs"]. Thus cDNA sequences are immediately useful for the identification of genes that are actually expressed as proteins, and they may even prove useful for studying protein structure and function if we get that far.

We think we can sequence a lot of cDNAs without slowing down the momentum of the mapping effort. This effort will probably catalyze a lot of activity in the community because it lends itself to independent participation by individuals in small labs. It's only when we try to collate, organize, and distribute the sequence data that we need a high level of coordination.

Bob Moyzis: Just as the genome mapping was divided by chromosome among various labs, the cDNA work can be distributed among different groups.

David Galas: Recently, people have found ways to make better libraries of cDNA clones, better in the sense that they more evenly represent the different mRNAs produced in the cell. These *normalized* cDNA libraries are trivial to produce in comparison with YAC libraries, and they're being generated in new cloning vectors designed to facilitate standard sequencing reactions. In six months to one year, we should have a bunch of good cDNA libraries being tested and sequenced.

As long as we have to sequence short stretches of DNA to make STSs, why not choose those short stretches from cDNAs rather than from some random set of DNA fragments?

We still don't know how easy it will be to map the cDNA sequences to specific chromosomes because not all cDNAs will be unique. On the other hand, preliminary data suggest that some STSs derived from human cDNAs will serve as STSs for other species, such as the mouse. Those cross-species STSs may help us enormously in comparing the mouse genome and the human genome.

Bob Moyzis: We should explain that different proteins and protein families share many similarities. Evolution is conservative and did not re-invent the wheel every time a protein with a new function appeared. Rather, gene duplication and rearrangement was used to produce novel proteins. So, short stretches of a cDNA sequence may appear in many different human genes. Regions of this kind make poor STS markers because they tag multiple sites.

It is still uncertain what the overall efficiency of producing STSs from cDNAs will be. If the goal is to tag genes, genomic DNA sequencing may be just as efficient. For example, from our work on chromosomes 5 and 16, random STS generation appears to be uncovering a significant fraction of coding regions. This is not unexpected since with over 100,000 genes and a sequencing window of 400 nucleotides, approximately one third of the sequences should contain a piece of a coding region.

Therefore, as we make the 30,000 STSs for the physical map of the genome, we are likely to find pieces of approximately 10,000 genes. The advantage of sequencing cDNAs rather than random fragments of genomic DNA will depend on what other information—and how much—can be obtained by this alternative approach.

David Galas: The initial purpose of using cDNA sequences as STS markers is to find unique landmarks for mapping that also fall within expressed genes. As long as a cDNA is unique, it is useful for that purpose. Additionally, cDNAs may be useful for determining whether genes are distributed evenly across the whole genome or are clustered together in certain regions.

Bob Moyzis: Yes, we'll learn something about gene density from cDNAs as well as genomic sequencing. We may also have to revise our estimate of the

number of genes in the genome. The present estimate of 50,000 to 100,000 genes is incredibly loose. It is based on theoretical arguments originally proposed by Haldane, which are now known to be based on false assumptions. The goal of the Human Genome Project is not only to map and sequence the genome, but ultimately to understand how it functions. At one extreme are people who believe that if we had the whole sequence of the human genome, we would be able to scan that sequence and find all of the protein coding regions, all the exons. But right now we don't know enough about the rules of the game to identify those regions unambiguously.

Neural-net programs like GRAIL, developed at Oak Ridge, are good, but not good enough. Some coding regions are very short, and these are hard to pick out directly from the sequence. Also, many genes have alternative sites for splicing out introns, so many messenger RNAs can be made from the same gene region. We can't yet predict from the DNA sequence which of those mRNAs are actually made. Consequently, at the opposite extreme are people who believe we must sequence every cDNA to unambiguously determine what protein is really being made from each particular gene. It seems clear that we'll need to sequence a lot of cDNAs in addition to sequencing the entire genome in order to learn the rules for finding genes.

Norton Zinder: I think the new emphasis on cDNAs may be very distracting to the goals of the Genome Project. cDNAs span about 10 percent of the genome, so if we sequence them with present technology, which is at least ten times more costly than what we are shooting for, then we will use up the whole genome budget on sequencing cDNAs.

David Galas: But the DOE is not proposing to sequence all the cDNAs but rather to sequence some cDNAs and use those sequences as a source of STS markers. We're putting a relatively small fraction of our resources, maybe a few percent, into finding out whether or not this approach will work. The cDNA effort will also mesh nicely with work on the mouse genome. The genetic-linkage map of the mouse is progressing very rapidly, in part because the mouse community is really behind the mapping effort. It is now apparent that genetic-linkage studies can be performed very efficiently in mice by crossing laboratory mice with wild mice from a different subspecies. Intraspecies crosses create offspring that are heterozygous for almost every genetic marker. In other words, each offspring carries two forms, or alleles, of almost every genetic marker. When those offspring are involved in controlled matings, a small number of successive generations of mice are sufficient to determine the genetic distances between many pairs of markers simultaneously.

The homology between the mouse genome and the human genome is very high. In fact, most of the cDNA sequences we find in humans will be present in mice with very little and sometimes no difference. People are often humbled at how closely we are related to mice. On the evolutionary scale, mice and humans have diverged a sufficiently short time ago that large stretches of mouse chromosomes can be matched up with corresponding stretches of human chromosomes. Therefore, the ordering of a bunch of cDNAs or a bunch of genetic markers will often be roughly the same on the mouse and on the human genomes. Consequently, studies of the mouse genome may provide some shortcuts for mapping certain regions of the human genome.

The NIH already has a program on mice, whereas the DOE genome program is largely focused on humans—and mostly on physical mapping. The coupling of the DOE genome program to the mouse project will most likely come through the cDNA work. We already have a huge mouse facility at the Oak Ridge Laboratory. It is the second largest such facility, the largest being at the Jackson Labs at Bar Harbor. We need to make those facilities available to the Genome Project, and we've recently funded a project at Oak Ridge.

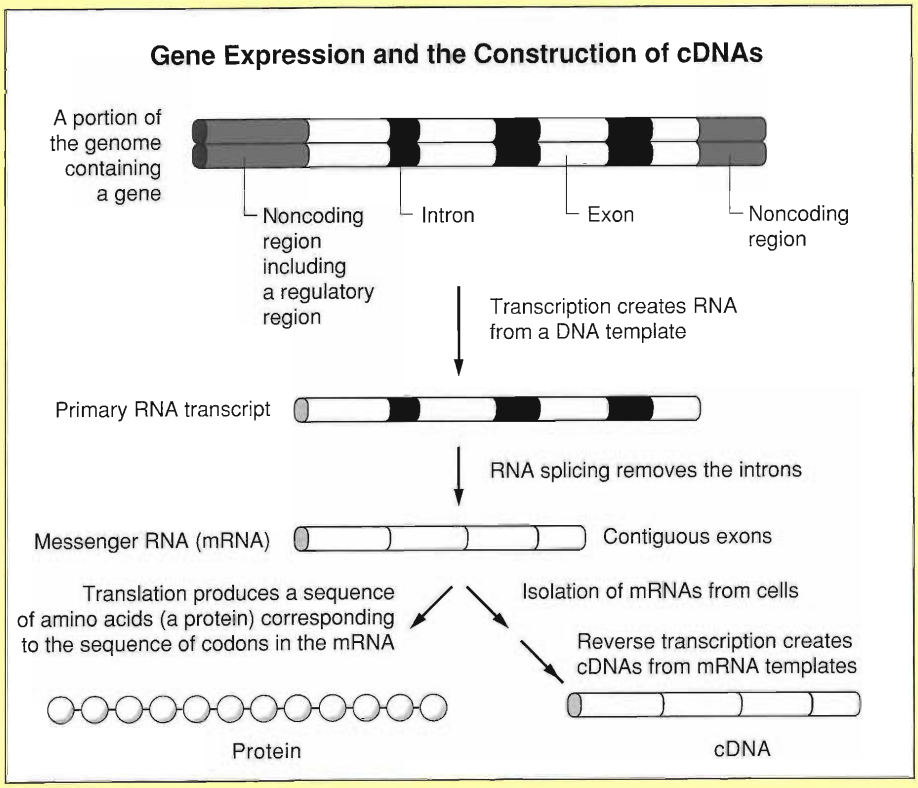
Maynard Olson: The reason the cDNA sequencing proposals are worrisome is that they sound so good. We are all interested in discovering new human genes by any method that gives us solid information, and the sequencing of cDNAs appears to do that. But you have to ask yourself: Would I support a Human Genome Project whose principle goal was to make a catalog of cDNA sequences? For me, the answer is no. Such a catalog of cDNAs will be of exceedingly low quality. We have learned that the route from RNA to cDNA to protein is ragged around the edges. It involves RNA editing, errors in reverse transcription of RNA to cDNA, and so on.

Biology only gets more complicated as you get away from the genome. At the DNA level, the genome is analogous to a relatively simple kind of computer disk file. But RNA, as we've become increasingly aware, is an extremely complicated molecule. Because it is single-stranded, it can fold up in complex ways, which, in turn, affect its function.

By the time you reach the complicated structures of proteins, you've got the subtle complexity that makes biology possible. The Genome Project has to

cDNAs and Expressed Genes

Copy DNAs, or cDNAs, are being synthesized, cloned, and sequenced as a source of STSs, unique landmarks for the physical map of the human genome. A cDNA is a copy of the protein-coding regions (exons) of a gene. It is not made directly from DNA isolated from the genome but rather, as shown in the figure, from the messenger RNA, the template that is translated into a protein. These templates are valuable because, unlike genomic DNA, each mRNA is a continuous stretch of protein-coding nucleotides. Moreover, the existence of an mRNA is proof that the corresponding protein-coding gene is an active, or expressed, gene.



cDNAs are synthesized in vitro. First, mRNAs are isolated from a population of tissue-specific cells. The isolated mRNAs represent only those genes that are being expressed in those particular cells. Each mRNA serves as a template in the synthesis of a complementary strand of DNA—the cDNA. The process of transcribing RNA into DNA, known as reverse transcription, is catalyzed by reverse transcriptase, an enzyme isolated from retroviruses, namely, RNA tumor viruses. The synthesized cDNAs are often shorter than the mRNA templates because of various processes that either degrade the mRNA or result in incomplete transcription. (Note that reverse transcriptase is not made by human cells. However, retroviruses, such as HIV, carry reverse transcriptase with them when they enter a host cell. The enzyme converts the viral RNA genome to DNA, which is then permanently incorporated into the genome of the host cell.)

After being synthesized in vitro, cDNAs are cloned. Cloned cDNAs have long been used for two purposes. First, cDNA libraries (random collections of cloned cDNAs) are used as sources of probes to identify the location of protein-coding regions in cloned fragments of genomic DNA. Second, particular mRNAs are isolated, converted to cDNAs, cloned, and then sequenced to determine the amino-acid sequence of the protein specified by the corresponding protein-coding gene.

The new emphasis is on sequencing short sections of cDNAs. If such a sequence is unique, it can be developed into a special kind of STS—one that is not only a unique, detectable landmark on the physical map of the genome but is also known to lie within an expressed gene. Furthermore, the cDNA sequence data provides some information about the protein encoded by the corresponding gene. ■

be directed at the DNA because DNA is the Achilles heel of the cell. It's the one thing we have a chance of understanding. Higher-order phenomena of alternative RNA splicing, reverse-transcription artifacts, and complicated gene families will be invisible in the rough cDNA sequences, so a catalog of cDNA sequences would be a very frustrating thing to work with. The first thing you would want to do with such a catalogue would be to map all the cDNAs onto the physical map to determine their chromosomal location. If you had several seemingly different cDNAs all mapped to the same place, you would have to look again and decide whether or not they really were different.

Lee Hood: A number of U.S. groups are sequencing cDNAs. This method will allow one to readily identify interesting genes, but each cDNA library will allow one to identify only the small subset of genes that are abundantly expressed in a particular cell type.

Many important genes cannot be obtained by this approach because of the rarity of their mRNAs. Polymorphisms and multigene families may also be challenging to decipher. I believe there is merit in sequencing both genomic DNA and cDNAs. The issue of whether fragments of DNA sequence can be patented is obviously very controversial.

Bob Moyzis: The added value of most cDNAs comes once they are mapped. Obtaining the sequence of the cDNA is the most trivial part of the process. Its difficult to see the basis for patenting cDNA sequences.

Lee Hood: In contrast, the large-scale sequencing of interesting regions in the genome has a guaranteed payoff. At Caltech we're sequencing the immune receptor loci, which will very likely lead to a much deeper understanding of autoimmune diseases and bring big biomedical and economic payoffs. If you focus sequencing efforts on DNA from the genome, you can direct those efforts to interesting regions. If you sequence cDNAs at random, it's the luck of the draw. Sequencing cDNAs is an inexpensive way of generating STSs to do physical maps. In the U.S., most scientists propose not to sequence the whole length of each cDNA but enough to generate unique genetic markers.

By the time you reach the complicated structures of proteins, you've got the subtle complexity that makes biology possible. The Genome Project has to be directed at the DNA because DNA is the Achilles heel of the cell. It's the one thing we have a chance of understanding.

Norton Zinder: My own guess is that technological breakthroughs will make it easier to blindly sequence the entire genome than to pick out specific regions using cDNAs and then go back and sequence those regions.

Maynard Olson: We had a similar debate about whether to sequence 10 percent of the genome or the whole genome. These debates become meaningless when we think more ambitiously about developing new technology. We're interested in seeing an order-of-magnitude reduction in sequencing costs. Now we are in a gray zone, and some people argue that the cost of sequencing is too high to justify sequencing the human genome but acceptably inexpensive to sequence all the cDNAs. If we are in this situation, we're unlikely to stay there.

Ultimately, we will want to sequence both the human genome and the cDNAs. In fact, I think that we're going to want to sequence both the human and the mouse genomes. Though we probably won't be able to find the exons by looking at the human DNA sequence alone, the way to find them is not to have some mishmash of cDNA sequences. We need to sequence the genomes of the human and the mouse, and maybe some other organisms, and then place them side by side and do comparisons. We should be putting more energy into the kind of technology development that would make that feasible.

If you want to study in detail the expression of a particular gene, you're going to have to do a lot of difficult experiments. We never claimed that lining up the mouse and human genome sequences would settle all the issues about which genes are expressed and which are not. But ask somebody who is trying to understand a particular gene whether he would like to compete in a situation where another lab had access to *both* the mouse and the human genomic sequence for the gene of interest while all he had was a bunch of cDNA clones. Then, you'd see more enthusiasm for mapping and a little less enthusiasm for cataloging cDNA sequences.

Bob Moyzis: If you want to understand how a gene works, when and where it gets expressed, and so on, you're never going to find that out from the cDNAs.

The Five-Year Goals

Bob Moyzis: I would guess that by the end of the Project's first five years, we will have some semblance of the high-resolution linkage maps, and, for some regions of some of the chromosomes, we will have reasonable physical maps. However, unless the effort on large-scale mapping increases, I don't think we will be able to complete the high-resolution physical maps on schedule—low-resolution contig maps, perhaps, but not the sets of closely spaced unique landmarks to go with the contig maps.

Maynard Olson: The prospect for meeting the Project's five-year goals doesn't look great at the moment, but that shouldn't be a matter of excessive concern. Right now we don't have enough mappers, in part because most molecular biologists are not trained in, nor are they necessarily good at, the analytical and technical skills required for the task. I think we will eventually recruit people who are not now working with DNA, and then momentum will build and the job will get done fairly quickly—although probably later than the famous five-year plan says. But remember, we asked for a $1-billion five-year plan—we're getting the half-price version. There will be a lag before new recruits enter the mapping effort, and we're in the lag phase now, but I predict that the maps will get done.

David Galas: At the moment physical mapping is perceived to be technology-limited. We are getting a lot of good mapping information, but the process is slow and tedious. However, it's foolish to think that the technology won't improve, and it may improve dramatically because a lot of innovation is still going on.

Lee Hood

So far the NIH leaders have been far too timid about making decisions that allow the money to be spent . . . appropriately. That has to change. If it doesn't . . . we're not going to come close to meeting the five-year goals.

Bob Moyzis: Our recent success at Los Alamos in producing chimera-free, chromosome-specific YAC libraries is an innovation that will have a significant impact on our own mapping effort as well as on the efforts of other genome centers. And we can expect to see other improvements in mapping technology. However, even now the technology is good enough, I feel, to complete the maps in five years.

The last ten years have seen major advances in technology development, such as YAC cloning. The major problem in achieving our goals is, as Maynard has mentioned, the lack of funding directed specifically at physical mapping and the lack of individuals who are truly interested in generating the maps.

Most of the people who would like to be funded by the Genome Project would rather try to improve mapping technology than make the maps. But if we can get the maps in five years with current technology, why spend the Project's money on technology improvements that will take five years to develop? It may be that in a hundred years we will be able to map a genome in a day, but I don't want to wait that long if the goal can be achieved in five years with current technology. I'd rather see technology-development money spent on the real bottlenecks, namely, sequencing and information management, analysis, and distribution.

Lee Hood: The Europeans seem more willing to give appropriate support to big projects. I'm not advocating creating a network of thirty-five laboratories to sequence one particular yeast chromosome, but that's what they did in Europe, and all 300,000 base pairs of the chromosome got sequenced. We haven't done a comparable project in the United States.

It's also obvious from the large investments in automation being made at CEPH [Centre d'Etude du Polymorphisme Humain] that the French government is willing to put a lot of money into carrying out very-large-scale linkage mapping. The CEPH scientists have built robots to make Southern blots of DNA from their repository of family cell lines. They will give those blots as

well as probes for various DNA markers to about thirty labs and ask them to identify which forms of the markers are present in the blots.

The project is a big one, and it's attractive to the participating laboratories because they are going to get paid more than it's going to cost them to do the work. CEPH will then stand at least a chance of putting together a very good linkage map in a reasonably short time. You don't see that kind of commitment at any of the U.S. centers that are carrying out linkage mapping.

What the NIH is tending to do with its genome centers is to nickel-and-dime them to death. The centers put in reasonably ambitious proposals, and the proposals come back with cuts in equipment, in computers, in technicians, and so on. Most important, virtually all of the funding for technology development was cut. Much of the NIH budget is being spread over small projects that won't amount to much.

To turn this around, we need determined leaders in both the DOE and the NIH. They must make a commitment to spend money in ways that will get the objectives done. So far the NIH leaders have been far too timid about making decisions that allow the money to be spent programmatically and appropriately. That has to change. If it doesn't, I would agree with Bob: We're not going to come close to meeting the five-year goals.

Mapping a chromosome is an enormous task, and we in the United States are going to have to come to grips with that fact. If we're going to be stingy about supporting the people who have already taken on such big projects, we're not going to encourage other people to take them on.

The Five-Year Goals of the U.S. Human Genome Project

Genetic Map Complete a fully connected human genetic map with markers spaced an average of 2 to 5 centimorgans apart. Identify each marker by an STS.

Physical Map Assemble STS maps of all human chromosomes with the goal of having markers spaced at approximately 100,000 base pair intervals. Generate overlapping sets of cloned DNA or closely spaced, unambiguously ordered markers with continuity over lengths of 2 million base pairs for large parts of the human genome.

DNA Sequencing Improve current methods and/or develop new methods for DNA sequencing that will allow large-scale sequencing of DNA at a cost of 50 cents per base pair. Determine the sequence of an aggregate of 10 million base pairs of human DNA in large continuous stretches in the course of technology development and validation.

Model Organisms Prepare a genetic map of the mouse genome based on DNA markers. Start physical mapping on one or two chromosomes. Sequence an aggregate of about 20 million base pairs of DNA from a variety of model organisms, focusing on stretches that are 1 million base pairs long, in the course of the development and validation of new and/or improved DNA-sequencing technology.

Informatics: Data Collection and Analysis Develop effective software and database designs to support large-scale mapping and sequencing projects. Create database tools that provide easy access to up-to-date physical mapping, genetic mapping, chromosome mapping, and sequencing information and allow ready comparison of the data in these several data sets. Develop algorithms and analytical tools to interpret genomic information.

Ethical, Legal, and Social Considerations Develop programs addressed at understanding the ethical, legal, and social implications of the Human Genome Project. Identify and define the major issues and develop initial policy options to address them.

Research Training Support research training of pre- and post-doctoral fellows starting in FY 1990. Increase the numbers of trainees supported until a steady state of about 600 per year is reached by the fifth year. Examine the need for other types of research training in the next year.

Technology Development Support innovative and high-risk technological developments as well as improvements in current technology to meet the needs of the Genome Project as a whole.

Technology Transfer Enhance the already close working relationship with industry. Encourage and facilitate the transfer of technologies and of medically important information to the medical community.

(From *Understanding Our Genetic Inheritance. The U.S. Human Genome Project: The First Five Years, FY 1991–1995.* NIH Publication No. 90–1590, April 1990.)